Optimal transport: discretization and algorithms

Quentin Merigot, Boris Thibert

OPTIMAL TRANSPORT: DISCRETIZATION AND ALGORITHMS

QUENTIN MÉRIGOT AND BORIS THIBERT

Abstract. This chapter describes techniques for the numerical resolution of optimal transport problems. We will consider several discretizations of these problems, and we will put a strong focus on the mathematical analysis of the algorithms to solve the discretized problems. We will describe in detail the following discretizations and corresponding algorithms: the assignment problem and Bertsekas auction's algorithm; the entropic regularization and Sinkhorn-Knopp's algorithm; semi-discrete optimal transport and Oliker-Prussner or damped Newton's algorithm, and finally semi-discrete entropic regularization. Our presentation highlights the similarity between these algorithms and their connection with the theory of Kantorovich duality.

Contents

1. Introduction

The problem of optimal transport, introduced by Gaspard Monge in 1871 [76], was motivated by military applications. The goal was to find the most economical way to transport a certain amount of sand from a quarry to a construction site. The source and target distributions of sand are seen as probability measures, denoted μ and ν, and $c(x, y)$ denotes the cost of transporting a grain of sand from the position x to the position y, and the goal is to solve the non-convex optimization problem

$$(\mathrm{MP}) = \min_{T_{\#}\mu=\nu} \int c(x, T(x))\mathrm{d}\mu, \tag{1.1}$$

where $T_{\#}\mu = \nu$ means that ν is the push-forward of μ under the transport map T. The modern theory of optimal transport has been initiated by Lenoid Kantorovich in the 1940s, via a convex relaxation of Monge's problem. Given two probability measures μ and ν, it consists in minimizing

$$(\mathrm{KP}) = \min_{\gamma\in\Gamma(\mu,\nu)} \int c(x, y)\mathrm{d}\gamma(x, y), \tag{1.2}$$

over the set $\Gamma(\mu, \nu)$ of *transport plans*[1] between μ and ν. Kantorovich's theory has been used and revisited by many authors from the 1980s, allowing a complete solution to Monge's problem in particular for $c(x, y) = \|x - y\|^p$. Since then, optimal transport has been connected to various domains of mathematics (geometry, probabilities, partial differential equations) but also to more applied domains. Current applications of optimal transport include machine learning [83], computer graphics [87], quantum chemistry [22, 32], fluid dynamics [20, 39, 73], optics [80, 26, 99, 24], economy [49], statistics [27, 30, 61]. The selection of citation above is certainly quite arbitrary, as optimal transport is now more than ever a vivid topic, with more than several hundreds (perhaps even thousands) of articles published every year and containing the words «optimal transport».

There exist many books on the theory of optimal transport, e.g. by Rachev-Rüschendorf [84, 85], by Villani [97, 98] and by Santambrogio [88]. However, there exist fewer books dealing with the numerical aspects, by Galichon [49], by Cuturi-Peyré [83] and one chapter of Santambrogio [88]. The books by Galichon and Cuturi-Peyré are targeted toward applications (in economy and machine learning, respectively) and do not deal in full detail with the mathematical analysis of algorithms for optimal transport. In this chapter, we concentrate on numerical methods for optimal transport relying on Kantorovich duality. Our aim in particular is to provide a self-contained mathematical analysis of several popular algorithms to solve the discretized optimal transport problems.

Kantorovich duality. In the 2000s, the theory of optimal transport was already mature and was used within mathematics, but also in theoretical physics or in economy. However, numerical applications were essentially

[1] A probability measure γ is a transport plan between μ and ν if its marginals are μ and ν.

limited to one-dimensional problems because of the prohibitive cost of existing algorithms for higher dimensional problems, whose complexity was in general more than quadratic in the size of the data. Numerous numerical methods have been introduced since then. Most of them rely the dual problem associated to Kantorovich's problem (1.2), namely

$$(\mathrm{DP}) = \max_{\varphi \ominus \psi \leqslant c} \int \varphi \mathrm{d}\mu - \int \psi \mathrm{d}\nu, \tag{1.3}$$

where the maximum is taken over pairs (φ, ψ) of functions satisfying $\varphi \ominus \psi \leqslant c$, meaning that $\varphi(x) - \psi(y) \leqslant c(x, y)$ for all x, y. Equivalently, the dual problem can be written as the unconstrained maximization problem

$$(\mathrm{DP}) = \max_{\psi} \mathcal{K}(\psi) \quad \text{where} \quad \mathcal{K}(\psi) = \int \psi^c \mathrm{d}\mu - \int \psi \mathrm{d}\nu \tag{1.4}$$

and $\psi^c(x) := \min_y c(x, y) + \psi(y)$ is the c-transform of ψ, a notion closely related to the Legendre-Fenchel transform in convex analysis. The function $\mathcal{K}$ is called the *Kantorovitch functional*. Kantorovich's duality theorem asserts that the values of (1.2) and (1.3) (or (1.4)) agree under mild assumptions.

Overview of numerical methods. We now briefly review the most used numerical methods for optimal transport. Note that there is no «free lunch» in the sense that there exists no method able to deal efficiently with arbitrary cost function; the computational complexity of most methods depend on the complexity of computing c-transforms or smoothed c-transforms. In this overview, we skip linear programming methods such as the network simplex, for which we refer to [83].

A. Assignment problem. When the two measures are uniformly supported on two finite sets with cardinal N, the optimal transport problem coincides with the *assignment problem*, described in [21]. The assignment problem can be solved using various techniques, but in this chapter we will concentrate on a dual ascent method called Bertsekas' auction algorithm [16], whose complexity is also $\mathrm{O}(N^3)$, but which is very simple to implement and analyze. In §3.2 we note that *the complexity can be improved when it is possible to compute discrete c-transforms efficiently*, namely

$$\psi^c(x_i) = \min_j c(x_i, y_j) + \psi(y_j). \tag{1.5}$$

B. Entropic regularization. In this approach, one does not solve the original optimal transport problem (1.2) exactly, but instead replaces it with a regularized problem involving the entropy of the transport plan. In the discrete case, it consists in minimizing

$$\min_{\gamma} \sum_{i,j} \gamma_{i,j} c(x_i, y_j) + \eta \sum_{i,j} h(\gamma_{i,j}), \tag{1.6}$$

where $h(r) = r(\log r - 1)$ and $\eta > 0$ is a small parameter, under the constraints

$$\forall i, \ \sum_j \gamma_{i,j} = \mu_i, \qquad \forall j, \ \sum_i \gamma_{i,j} = \nu_j \tag{1.7}$$

This idea has been introduced in the field of optimal transport by Galichon and Salanié [50] and by Cuturi [35], see [83, Remark 4.5] for a brief historical account. Adding the entropy of the transport plan makes the problem (1.6) strongly convex and smooth. The dual problem can be solved efficiently using Sinkhorn-Knopp's algorithm, which involves computing repeatedly the *smoothed c-transform*

$$\psi^{c,\eta}(x_i) = \eta \log(\mu_i) - \eta \log \left(\sum_j e^{\frac{1}{\eta}(-c(x_i,y_j)-\psi(y_j))} \right). \tag{1.8}$$

Sinkhorn-Knopp's algorithm can be very efficient, *provided that the smoothed c-transform can be computed efficiently (e.g. in near-linear time).*

C. Distance costs. When the cost c satisfies the triangle inequality, the dual problem (1.3) can be further simplified:

$$\max_{\mathrm{Lip}_c(\psi)\leqslant 1} \int \psi \mathrm{d}\mu - \int \psi \mathrm{d}\nu, \tag{1.9}$$

where the maximum is taken over functions satisfying $|\psi(x) - \psi(y)| \leqslant c(x,y)$ for all x, y. The equality between the values of (1.2) and (1.9) is called Kantorovich-Rubinstein's theorem. This leads to very efficient algorithms when the 1-Lipschitz constraint can be enforced using only *local* information, thus reducing the number of constraints. This is possible when the space is discrete and the distance is induced by a graph, or when c is the Euclidean norm or more generally a Riemannian metric. In the latter case, the maximum in (1.9) can be replaced by a supremum over $\mathcal{C}^1$ functions ψ satisfying $\|\nabla\psi\|_\infty \leqslant 1$ [94, 9]. Note that *the case of distance costs is particularly easy because the c-transform of a* 1*-Lipschitz function is trivial:* $\psi^c = -\psi$.

D. Monge-Ampère equation. When the cost is the Euclidean scalar product, $c(x,y) = -\langle x|y\rangle$, the dual problem (1.3) can be reformulated as

$$\max_{\psi} - \int \psi^* \mathrm{d}\mu - \int \psi \mathrm{d}\nu, \tag{1.10}$$

where $\psi^*(x) = \max_y \langle x|y\rangle - \psi(y)$ is the Legendre-Fenchel transform of ψ. If the maximizer ψ is smooth and strongly convex and μ, ν are probability densities, the optimality condition associated to the dual problem is the *Monge-Ampère equation,*

$$\begin{cases} \mu(\nabla\psi(y)) \det(\mathrm{D}^2\psi(y)) = \nu(y), \\ \nabla\psi(\mathrm{spt}(\nu)) \subseteq \mathrm{spt}(\mu). \end{cases} \tag{1.11}$$

Note the non-standard boundary conditions appearing on the second line of the equation. The first methods able to deal with these boundary conditions use a "wide-stencil" finite difference discretization [46, 13, 12]. These methods are able to solve optimal transport problems provided that the maximizer of (1.10) is a viscosity solution to the Monge-Ampère equation (1.11), imposing restrictions on its regularity. For the Monge-Ampère equation with Dirichlet conditions, we refer to the recent survey by Neilan, Salgado and Zhang [77].

E. Semi-discrete formulation. The semi-discrete formulation of optimal transport involves a source measure that is a probability density μ and a target measure ν which is finitely supported, i.e. $\nu = \sum_i \nu_i \delta_{y_i}$. It was introduced by Cullen in 1984 [34], without reference to optimal transport, and much refined since then [6, 70, 38, 55, 64, 68, 65]. In this setting, the dual problem (1.4) amounts to maximizing the Kantorovitch functional given by

$$\mathcal{K}(\psi) = \int \psi^c \mathrm{d}\mu - \int \psi \mathrm{d}\nu = \sum_i \int_{\mathrm{Lag}_{y_i}(\psi)} c(x, y_i) + \psi(y_i) \mathrm{d}\mu(x) - \int \psi \mathrm{d}\nu, \tag{1.12}$$

where the Laguerre cells are defined by

$$\mathrm{Lag}_{y_i}(\psi) = \{x \mid \forall j, c(x, y_i) + \psi(y_i) \leqslant c(x, y_j) + \psi(y_j)\}. \tag{1.13}$$

The optimality condition for (1.12) is the following non-linear system of equations,

$$\forall i, \mu(\mathrm{Lag}_{y_i}(\psi)) = \nu_i. \tag{1.14}$$

In the case $c(x, y) = -\langle x|y\rangle$, this system of equations can see as a weak formulation (in the sense of Alexandrov, see [57, Chapter 1]) of the Monge-Ampère equation (1.11). This "semi-discrete" approach can also be used to solve Monge-Ampère equations with Dirichlet boundary conditions, and has been originally introduced for this purpose [79, 75]. *Again, the possibility to solve* (1.14) *efficiently requires one to be able to compute the Laguerre tessellation* (1.13)*, and thus the c-transform, efficiently.*

F. Dynamic formulation. This formulation relies on the dynamic formulation of optimal transport, which holds when the cost is $c(x, y) = \|x - y\|^2$ on $\mathbb{R}^d$ (or more generally induced by a Riemannian metric), and is known as the *Benamou-Brenier formulation*:

$$\min_{(\rho,v)} \int_0^1 \int_{\mathbb{R}^d} \rho_t \|v_t\|^2 \mathrm{d}x\mathrm{d}t \quad \text{with} \quad \begin{cases} \partial_t \rho_t + \mathrm{div}(\rho_t v_t) = 0, \\ \rho_0 = \mu, \rho_1 = \nu. \end{cases}$$

Introducing the momentum $m_t = \rho_t v_t$, the problem can be rewritten as

$$\min_{(\rho_t,m_t)} \int_0^1 \int_{\mathbb{R}^d} \frac{\|m_t\|^2}{\rho_t} \mathrm{d}x\mathrm{d}t \quad \text{with} \quad \begin{cases} \partial_t \rho_t + \mathrm{div}(m_t) = 0, \\ \rho_0 = \mu, \rho_1 = \nu. \end{cases}$$

This optimization problem can be discretized using finite elements [7], finite differences [81] or finite volumes [44], and the discrete problem is then usually solved using a primal-dual augmented Lagrangian method [7] (see also [81, 60]). In practice, the convergence is very costly in terms of number of iterations; note also that each iteration requires the resolution of a $(d + 1)$-dimensional Poisson problem to project on the admissible set $\{(\rho, m) \mid \partial_t \rho + \mathrm{div}(m) = 0\}$. Another possibility is to use divergence-free wavelets [59]. One advantage of the Benamou-Brenier approach is that it is very flexible, easily allowing (Riemannian) cost functions [81], additional quadratic terms [8], penalization of congestion [23], partial transport [69, 31], etc. Finally, the convergence from the discretized problem to the continuous one is subtle and depends on the choice of the discretization, see [67, 28].

In this chapter, we will describe in detail the following discretizations for optimal transport and corresponding algorithms to solve the discretized problems : the assignment problem (A.) through Bertsekas auction's algorithm, the entropic regularization (B.) through Sinkhorn-Knopp's algorithm, semi-discrete optimal transport (E.) through Oliker-Prussner or Newton's methods. These algorithms share a common feature, in that they are all derived from Kantorovich duality. Some of them have been adapted to variants of optimal transport problems, such as multi-marginal optimal transport problems problems [82], barycenters with respect to optimal transport metrics [2], partial [25] and unbalanced optimal transport [31, 66], gradient flows in the Wasserstein space [62, 4], generated Jacobian equations [56, 95]. However, we consider these extensions to be out of the scope of this chapter.

2. Optimal transport theory

This part contains a self-contained introduction to the theory of optimal transport, putting a strong emphasis on Kantorovich Kantorovich duality. Kantorovich duality is at the heart of the most important theorems of optimal transport, such as Brenier and Gangbo-McCann's theorems on the existence and uniqueness of solution to Monge's problems and the stability of optimal transport plans and optimal transport maps maps. Kantorovich's duality is also used in all the numerical methods presented in this chapter.

Background on measure theory. In the following, we assume that X is a compact metric space, and we denote $\mathcal{C}^0(X)$ the space of continuous functions over X. We denote $\mathcal{M}(X)$ the space of *finite (Radon) measures* over X, identified with the set of continuous linear forms over $\mathcal{C}^0(X)$. Given $\varphi \in \mathcal{C}^0(X)$ and $\mu \in \mathcal{M}(X)$, we will often denote $\langle\varphi|\mu\rangle = \int_X \varphi \mathrm{d}\mu$. The spaces of *non-negative measures* and *probability measures* are defined by

$$\mathcal{M}^+(X) := \{\mu \in \mathcal{M}(X) \mid \mu \geqslant 0\},$$

$$\mathcal{P}(X) := \{\mu \in \mathcal{M}^+(X) \mid \mu(X) = 1\},$$

where $\mu \geqslant 0$ means $\langle\mu|\varphi\rangle \geqslant 0$ for all $\varphi \in \mathcal{C}^0(X, \mathbb{R}^+)$. The three spaces $\mathcal{M}(X)$, $\mathcal{M}^+(X)$ and $\mathcal{P}(X)$ are endowed with the weak topology induced by duality with $\mathcal{C}^0(X)$, namely $\mu_n \to \mu$ weakly if

$$\forall\varphi \in \mathcal{C}^0(X),\ \langle\varphi|\mu_n\rangle \xrightarrow{n\to\infty} \langle\varphi|\mu\rangle.$$

A point x belongs to the *support* of a non-negative measure μ iff for every $r > 0$ one has $\mu(\mathrm{B}(x,r)) > 0$. The support of μ is denoted $\mathrm{spt}(\mu)$. We recall that by Banach-Alaoglu theorem, the set of probability measures $\mathcal{P}(X)$ is weakly compact, a fact which will be useful to prove existence and convergence results in optimal transport.

Notation. Given two functions $\varphi \in \mathcal{C}^0(X)$ and $\psi \in \mathcal{C}^0(Y)$ we will define $\varphi \oplus \psi \in \mathcal{C}^0(X \times Y)$ by $\varphi \oplus \psi(x, y) = \varphi(x) + \psi(y)$. We define $\varphi \ominus \psi$ and $\varphi \otimes \psi$ similarly.

2.1. **The problems of Monge and Kantorovich.**

Monge's problem. Before introducing Monge's problem, we recall the definition of *push-forward* or *image* measure.

Definition 1 (Push-forward and transport map)**.** Let X, Y be compact metric spaces, $\mu \in \mathcal{M}(X)$ and $T : X \to Y$ be a measurable map. The *push-forward* of μ by T is the measure $T_{\#}\mu$ on Y defined by

$$\forall \varphi \in \mathcal{C}^0(Y), \langle \varphi | T_{\#}\mu \rangle := \langle \varphi \circ T | \mu \rangle,$$

or equivalently if for every Borel subset $B \subseteq Y$, $\mathrm{T}_{\#}\mu(B) = \mu(T^{-1}(B))$. A measurable map $T : X \to Y$ such that $T_{\#}\mu = \nu$ is also called a *transport map* between μ and ν.

Example 1. If $Y = \{y_1, \dots, y_n\}$, then $T_{\#}\mu = \sum_{1 \leqslant i \leqslant n} \mu(T^{-1}(\{y_i\}))\delta_{y_i}$.

Example 2. Assume that T is a $\mathcal{C}^1$ diffeomorphism between compact domains X, Y of $\mathbb{R}^d$, and assume also that the probability measures μ, ν have continuous densities ρ, σ with respect to the Lebesgue measure. Then,

$$\int_Y \varphi(y)\sigma(y)\mathrm{d}y = \int_X \varphi(T(x))\sigma(T(x)) \det(\mathrm{D}T(x))\mathrm{d}x.$$

Hence, T is a transport map between μ and ν iff

$$\forall \varphi \in \mathcal{C}^0(X), \int_X \varphi(T(x))\sigma(T(x)) \det(\mathrm{D}T(x))\mathrm{d}x = \int_X \varphi(T(x))\rho(x)\mathrm{d}x,$$

or equivalently if the (non-linear) Jacobian equation holds

$$\rho(x) = \sigma(T(x)) \det(\mathrm{D}T(x)).$$

Definition 2 (Monge's problem)**.** Consider two compact metric spaces X, Y, two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a *cost function* $c \in \mathcal{C}^0(X \times Y)$. *Monge's problem* is the following optimization problem

$$\text{(MP)} := \inf \left\{ \int_X c(x, T(x))\mathrm{d}\mu(x) \mid T : X \to Y \text{ and } T_{\#}\mu = \nu \right\} \tag{2.15}$$

Monge's problem exhibits several difficulties, one of which is that both the transport constraint ($T_{\#}\mu = \nu$) and the functional are non-convex. Note also that there might exist no transport map between μ and ν. For instance, if $\mu = \delta_x$ for some $x \in X$, then, $\mathrm{T}_{\#}\mu(B) = \mu(T^{-1}(B)) = \delta_{T(x)}$. In particular, if $\mathrm{card}(\mathrm{spt}(\nu)) > 1$, there exists no transport map between μ and ν.

Kantorovich's problem.

Definition 3 (Marginals)**.** The *marginals* of a measure γ on a product space $X \times Y$ are the measures $\Pi_{X\#}\gamma$ and $\Pi_{Y\#}\gamma$, where $\Pi_X : X \times Y \to X$ and $\Pi_Y : X \times Y \to Y$ are their projection maps.

Definition 4 (Transport plan)**.** A transport plan between two probability measures μ, ν on two metric spaces X and Y is a probability measure γ on the product space $X \times Y$ whose marginals are μ and ν. The space of transport plans is denoted $\Gamma(\mu, \nu)$, i.e.

$$\Gamma(\mu, \nu) = \{\gamma \in \mathcal{P}(X \times Y) \mid \Pi_{X\#}\gamma = \mu,\ \Pi_{Y\#}\gamma = \nu\}.$$

Note that $\Gamma(\mu, \nu)$ is a convex set.

Example 3 (Product measure). Note that the set of transport plans $\Gamma(\mu,\nu)$ is never empty, as it contains the measure $\mu \otimes \nu$.

Definition 5 (Kantorovich's problem)**.** Consider two compact metric spaces X, Y, two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a *cost function* $c \in \mathcal{C}^0(X \times Y)$. *Kantorovich's problem* is the following optimization problem

$$\text{(KP)} := \inf \left\{ \int_{X\times Y} c(x,y) \mathrm{d}\gamma(x,y) \mid \gamma \in \Gamma(\mu,\nu) \right\} \tag{2.16}$$

Remark 1. The infimum in Kantorovich's problem is less than the infimum in Monge's problem. Indeed, to any transport map T between μ and ν one can associate a transport plan, by letting $\gamma_T = (\mathrm{id}, T)_\#\mu$. One can easily check that $\Pi_{X\#}\gamma_T = \mu$ and $\Pi_{Y\#}\gamma_T = \nu$ so that $\gamma_T \in \Gamma(\mu,\nu)$ is a transport plan between μ and ν. Moreover, by the definition of push-forward,

$$\langle c|\gamma_T\rangle = \langle c|(\mathrm{id}, T)_\#\mu\rangle = \langle c \circ (\mathrm{id}, T)|\mu\rangle = \int_X c(x, T(x))\mathrm{d}\mu$$

thus showing that (KP) $\leqslant$ (MP).

Proposition 1. *Kantorovich's problem* (KP) *admits a minimizer.*

Proof. The definition of $\Pi_{X\#}\gamma = \mu$ can be expanded into

$$\forall \varphi \in \mathcal{C}^0(X), \langle \varphi \otimes 1|\gamma\rangle = \langle \varphi|\mu\rangle,$$

from which it is easy to see that the set $\Gamma(\mu,\nu)$ is weakly closed, and therefore weakly compact as a subset of $\mathcal{P}(X \times Y)$, which is weakly compact by Banach-Alaoglu's theorem. We conclude the existence proof by remarking that the functional that is minimized in (KP), namely $\mu \mapsto \langle c|\mu\rangle$, is weakly continuous by definition. □

2.2. **Kantorovich duality.**

Derivation of the dual problem. The primal Kantorovich problem (KP) can be reformulated by introducing Lagrange multipliers for the constraints. Namely, we use that for any $\gamma \in \mathcal{M}^+(X \times Y)$,

$$\sup_{\varphi \in \mathcal{C}^0(X)} -\langle \varphi \otimes 1|\gamma\rangle + \langle \varphi|\mu\rangle = \begin{cases} 0 & \text{if } \Pi_{X\#}\gamma = \mu \\ +\infty & \text{if not} \end{cases}$$

$$\sup_{\varphi \in \mathcal{C}^0(X)} \langle 1 \otimes \psi|\gamma\rangle - \langle \psi|\mu\rangle = \begin{cases} 0 & \text{if } \Pi_{X\#}\gamma = \mu \\ +\infty & \text{if not} \end{cases}$$

to deduce

$$\sup_{\varphi \in \mathcal{C}^0(X), \psi \in \mathcal{C}^0(Y)} \langle \varphi|\mu\rangle - \langle \psi|\nu\rangle - \langle \varphi \ominus \psi|\gamma\rangle = \begin{cases} 0 & \text{if } \gamma \in \Gamma(\mu,\nu) \\ +\infty & \text{if not.} \end{cases}$$

This leads to the following formulation of the Kantorovich problem

$$\text{(KP)} = \inf_{\gamma \in \mathcal{M}^+(X\times Y)} \sup_{(\varphi,\psi) \in \mathcal{C}^0(X)\times\mathcal{C}^0(Y)} \langle c - (\varphi \ominus \psi)|\gamma\rangle + \langle \varphi|\mu\rangle - \langle \psi|\nu\rangle$$

Kantorovich dual problem is simply obtained by inverting the infimum and the supremum:

$$\text{(DP)} := \sup_{\varphi,\psi} \inf_{\gamma \geqslant 0} \langle c - (\varphi \ominus \psi)|\gamma\rangle + \langle \varphi|\mu\rangle - \langle \psi|\nu\rangle.$$

Note that we will often omit the assumptions that $\gamma \in \mathcal{M}(X \times Y)$ and φ, ψ are continuous, when the context is clear. The dual problem can further be simplified by remarking that

$$\inf_{\gamma \geqslant 0} \langle c - \varphi \ominus \psi | \gamma \rangle = \begin{cases} 0 & \text{if } \varphi \ominus \psi \leqslant c \\ -\infty & \text{if not.} \end{cases}$$

Definition 6 (Kantorovich's dual problem)**.** Given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ with X, Y compact metric spaces and $c \in \mathcal{C}^0(X \times Y)$, we define Kantorovich's dual problem by

$$(\mathrm{DP}) = \sup \left\{ \int_X \varphi \mathrm{d}\mu - \int_Y \psi \mathrm{d}\nu \mid (\varphi, \psi) \in \mathcal{C}^0(X) \times \mathcal{C}^0(Y), \varphi \ominus \psi \leqslant c \right\} \tag{2.17}$$

Proposition 2. *Weak duality holds, i.e.* (KP) $\geqslant$ (DP)*.*

Proof. Given $(\varphi, \psi, \gamma) \in \mathcal{C}^0(X) \times \mathcal{C}^0(Y) \times \Gamma(\mu, \nu)$ satisfying the constraint $\varphi \ominus \psi \leqslant c$, one has

$$\langle \varphi | \mu \rangle - \langle \psi | \nu \rangle = \langle \varphi \ominus \psi | \gamma \rangle \leqslant \langle c | \gamma \rangle,$$

where we used $\gamma \in \Gamma(\mu, \nu)$ to get the equality and $\varphi \ominus \psi \leqslant c$ to get the inequality. As a conclusion,

$$(\mathrm{DP}) = \min_{\varphi \ominus \psi \leqslant c} \langle \varphi | \mu \rangle - \langle \psi | \nu \rangle \leqslant \max_{\gamma \in \Gamma(\mu, \nu)} \langle c | \gamma \rangle = (\mathrm{KP}) \qquad \square$$

Existence of solution for the dual problem. Kantorovich's dual problem (DP) consists in maximizing a concave (actually linear) functional under linear inequality constraints. It can also also easily be turned into an unconstrained minimization problem. The idea is quite simple: given a certain $\psi \in \mathcal{C}^0(Y)$, one wishes to select φ on X which is as large as possible (to maximize the term $\langle \varphi | \mu \rangle$ in (DP)) while satisfying the constraint $\varphi \ominus \psi \leqslant c$. This constraint can be rewritten as

$$\forall x \in X, \ \varphi(x) \leqslant \min_{y \in Y} c(x, y) + \psi(y).$$

The largest function φ satisfying it is $\varphi(x) = \min_{y \in Y} c(x, y) + \psi(y)$. Thus,

$$\begin{aligned} (\mathrm{KP}) &= \sup_{\varphi \ominus \psi \leqslant c} \langle \varphi | \mu \rangle - \langle \psi | \nu \rangle \\ &= \sup_{\psi \in \mathcal{C}^0(Y)} \int_X \left(\min_{y \in Y} c(x, y) + \psi(y) \right) \mathrm{d}\mu(x) - \int \psi(y) \mathrm{d}\nu(y). \end{aligned}$$

This idea is at the basis of many algorithms to solve discrete instances of optimal transport, but also useful in theory. It also suggests to introduce the notion of c-transform. .

Definition 7 (c-Transform)**.** The c-transform (resp. $\overline{c}$-transform) of a function $\psi : Y \to \mathbb{R} \cup \{+\infty\}$ (resp. $\varphi : X \to \mathbb{R} \cup \{+\infty\}$) is defined as

$$\psi^c : x \in X \mapsto \inf_{y \in Y} c(x, y) + \psi(y) \tag{2.18}$$

$$\varphi^{\overline{c}} : y \in Y \mapsto \sup_{x \in X} -c(x, y) + \varphi(x) \tag{2.19}$$

Thanks to this notion of c-transform, one can reformulate the dual problem (DP) as an unconstrained maximization problem:

$$(\mathrm{DP}) = \sup_{\psi \in \mathcal{C}^0(Y)} \int_X \psi^c \mathrm{d}\mu - \int_Y \psi \mathrm{d}\nu. \tag{2.20}$$

Remark 2 (c-concavity, $\overline{c}$-convexity and c-subdifferential). One can call a function φ on X *c-concave* if $\varphi = \psi^c$ for some $\psi : Y \to \mathbb{R} \cup \{+\infty\}$ on Y. Note that we use the word *concave* because ψ^c is defined through an infimum. Conversely, a function ψ on Y is called *$\overline{c}$-convex* if $\psi = \varphi^{\overline{c}}$ for some $\varphi : X \to \mathbb{R} \cup \{+\infty\}$. Note the asymetry between the two notions, which is due to the choice of the sign in the constraint in Kantorovich's problem: in the two equivalent formulations

$$(\mathrm{KP}) = \sup_{\varphi \oplus \psi \leqslant c} \langle \varphi | \mu \rangle + \langle \psi | \nu \rangle = \sup_{\varphi \ominus \psi \leqslant c} \langle \varphi | \mu \rangle - \langle \psi | \nu \rangle,$$

we chose the second one, involving two minus signs. This choice will make it easier to explain some of the algorithms we will present later in the chapter. The *c-subdifferential* of a function ψ on Y is a subset of $X \times Y$ defined by

$$\partial^c \psi := \{(x, y) \in X \times Y \mid \psi^c(x) - \psi(y) = c(x, y)\}, \tag{2.21}$$

while the *c-subdifferential* at a point y in Y is given by

$$\partial^c \psi(y) := \{x \in X, \quad (x, y) \in \partial^c \psi\}. \tag{2.22}$$

Remark 3 (Bilinear cost). When $c(x, y) = -\langle x | y \rangle$, a function is $\overline{c}$-convex if and only if it is convex, and $\varphi^{\overline{c}}$ is the Legendre-Fenchel transform of $-\varphi$.

Proposition 3 (Existence of dual potentials). (DP) *admits a maximizer, which one can assume to be of the form (φ, ψ) such that $\varphi = \psi^c$ and $\psi = \varphi^{\overline{c}}$.*

The existence of maximizers follows from the fact that a c-concave/$\overline{c}$-convex function has the same modulus of continuity as c.

(Recall that $\omega : \mathbb{R}^+ \to \mathbb{R}$ is a modulus of continuity of a function $f : Z \to \mathbb{R}$ on a metric space (Z, d_Z) if it satisfies $\lim_{t \to 0} \omega(t) = 0$ and for every $z, z' \in Z$, $|f(z) - f(z')| \leqslant \omega(\mathrm{d}_Z(z, z'))$.)

Lemma 4 (Properties of c-transforms). *Let $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ be a modulus of continuity for $c \in \mathcal{C}^0(X \times Y)$ for the distance*

$$\mathrm{d}_{X \times Y}((x, y), (x', y')) = \mathrm{d}_X(x, x') + \mathrm{d}_Y(y, y').$$

Then for every $\varphi \in \mathcal{C}^0(X)$ and every $\psi \in \mathcal{C}^0(Y)$,

- *$\varphi^{\overline{c}}$ and ψ^c also admits ω as modulus of continuity.*
- *$\psi^{c\overline{c}} \leqslant \psi$ and $\psi^{c\overline{c}c} = \psi^c$.*
- *$\varphi^{\overline{c}c} \geqslant \varphi$ and $\varphi^{\overline{c}c\overline{c}} = \varphi^{\overline{c}}$.*

Proof. Let us first prove the first point. Let $\psi \in \mathcal{C}^0(Y)$ and for $x \in X$, let $y_x \in Y$ be a point realizing the minimum in the definition of ψ^c. Then,

$$\psi^c(x') \leqslant c(x', y_x) + \psi(y_x) = \psi^c(x) + c(x', y_x) - c(x, y_x) \leqslant \psi^c(x) + \omega(\mathrm{d}_X(x, x')).$$

Exchanging the role of x and x' we get $|\psi^c(x') - \psi^c(x)| \leqslant \omega(\mathrm{d}_X(x, x'))$ as desired. The proof that $\varphi^{\overline{c}}$ has the ω as modulus of continuity is similar. We prove now the second point. By definition, one has

$$\psi^{c\overline{c}}(y) = \max_{x \in X} \left(-c(x, y) + \min_{\tilde{y} \in Y} c(x, \tilde{y}) + \psi(\tilde{y}) \right).$$

By taking $\tilde{y} = y$, one gets $\psi^{c\bar{c}}(y) \leqslant \psi(y)$. Again, by definition, we have

$$\psi^{c\bar{c}c}(x) = \min_{y \in Y} \left(c(x, y) + \max_{\tilde{x} \in X} \left(-c(\tilde{x}, y) + \min_{\tilde{y} \in Y} c(\tilde{x}, \tilde{y}) + \psi(\tilde{y}) \right) \right).$$

By taking $\tilde{x} = x$, one gets $\psi^{c\bar{c}c}(x) \geqslant \psi^c(x)$, while taking $\tilde{y} = y$ gives us $\psi^{c\bar{c}c}(x) \leqslant \psi^c(x)$. The last point is obtained similarly. □

Proof of Proposition 3. Let (φ_n, ψ_n) be a maximizing sequence for (DP), i.e. $\varphi_n \ominus \psi_n \leqslant c$ and $\lim_{n \to +\infty} \langle \varphi_n | \mu \rangle - \langle \psi_n | \nu \rangle = \text{(DP)}$. Define $\hat{\varphi}_n = \psi_n^c$ and $\hat{\psi}_n = \hat{\varphi}_n{}^{\bar{c}}$. Then $\hat{\varphi}_n \ominus \hat{\psi}_n \leqslant c$, $\varphi_n \leqslant \hat{\varphi}_n$ and $\psi_n \geqslant \hat{\psi}_n$, which implies

$$\langle \varphi_n | \mu \rangle - \langle \psi_n | \nu \rangle \leqslant -\langle \psi_n | \nu \rangle \leqslant \langle \hat{\varphi}_n | \mu \rangle - \langle \hat{\psi}_n | \nu \rangle,$$

implying that $(\hat{\varphi}_n, \hat{\psi}_n)$ is also a maximizing sequence. Our goal is now to show that this sequence admits a converging subsequence. We first note that we can assume that $\hat{\varphi}_n(x_0) = 0$ for all n, where x_0 is a given point in X: if this is not the case, we replace $(\hat{\varphi}_n, \hat{\psi}_n)$ by $(\hat{\varphi}_n - \hat{\varphi}_n(x_0), \hat{\psi}_n + \hat{\varphi}_n(x_0)))$, which is also admissible and has the same dual value. In addition, by Lemma 4, the sequences $(\hat{\varphi}_n)_n$ and $(\hat{\psi}_n)_n$ are equicontinuous. By Arzelà-Ascoli's theorem, we deduce that they admit subsequences converging respectively to $\varphi \in \mathcal{C}^0(x)$ and $\psi \in \mathcal{C}^0(Y)$, which are then maximizers for (DP). □

Strong duality and stability of optimal transport plans. We will prove strong duality first in the case where μ, ν are finitely supported, and will then use a density argument to deduce the general case. As a byproduct of this theorem, we get a stability result for optimal transport plans (i.e. a limit of optimal transport plans is also optimal).

Theorem 5 (Strong duality). *Let X, Y be compact metric spaces and $c \in \mathcal{C}^0(X \times Y)$. Then the maximum is attained in* (DP) *and* (KP) = (DP).

Corollary 6 (Support of OT plans). *Let ψ be a maximizer of* (2.20) *and $\gamma \in \Gamma(\mu, \nu)$ a transport plan. Then the two assertions are equivalent*

- *γ is an optimal transport plan*
- $\text{spt}(\gamma) \subset \partial^c \psi := \{(x, y) \in X \times Y \mid \psi^c(x) - \psi(y) = c(x, y)\}$.

As a consequence of Kantorovich duality, we can prove stability of optimal transport plans and optimal transport maps.

Theorem 7 (Stability of OT plans). *Let X, Y be compact metric spaces and let $c \in \mathcal{C}^0(X \times Y)$. Consider $(\mu_k)_{k \in \mathbb{N}}$ and $(\nu_k)_{k \in \mathbb{N}}$ in $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ converging weakly to μ and ν respectively.*

- *If $\gamma_k \in \Gamma(\mu_k, \nu_k)$ is optimal then, up to subsequences, (γ_k) converges weakly to an optimal transport plan $\gamma \in \Gamma(\mu, \nu)$.*
- *Let (φ_k, ψ_k) be optimal Kantorovich potentials in the dual problem between μ_k and ν_k, satisfying $\psi_k = \varphi_k^{\bar{c}}$ and $\varphi_k = \psi_k^c$. Given a point $x_0 \in X$, define $\tilde{\psi}_k = \psi_k - \psi_k(x_0)$ and $\tilde{\varphi}_k = \varphi_k + \psi_k(x_0)$. Then, up to subsequences, $(\tilde{\psi}_k, \tilde{\varphi}_k)$ converges uniformly to (φ, ψ) a maximizing pair for* (DP) *satisfying $\varphi = \psi^c$ and $\psi = \varphi^{\bar{c}}$.*

The proof of Theorem 5 relies on a simple reformulation of strong duality – similar to the Karush-Kuhn-Tucker optimality conditions for optimization problems with inequality constraints:

Proposition 8. *Let* $\gamma \in \Gamma(\mu, \nu)$ *and let* $(\varphi, \psi) \in \mathcal{C}^0(X) \times \mathcal{C}^0(Y)$ *such that* $\varphi \ominus \psi \leqslant c$*. Then, the following statements are equivalent:*

- $\varphi \ominus \psi = c$ γ*-a.e.*
- γ *minimizes* (KP)*,* (φ, ψ) *maximizes* (DP) *and* (KP) = (DP)*.*

Proof. Assume that $\varphi \ominus \psi = c$ γ-a.e. Then,

$$(\mathrm{KP}) \leqslant \langle c|\gamma\rangle = \langle \varphi \ominus \psi|\gamma\rangle = \langle \varphi|\mu\rangle - \langle \psi|\nu\rangle \leqslant (\mathrm{DP})$$

Since in addition (KP) $\geqslant$ (DP), all inequalities are equalities, which implies that (KP) = (DP), γ miminizes (KP) and (φ, ψ) maximizes (DP). Conversely, if (KP) = (DP), γ miminizes (KP) and (φ, ψ) maximizes (DP), then

$$\langle \varphi|\mu\rangle - \langle \psi|\nu\rangle = (\mathrm{DP}) = (\mathrm{KP}) = \langle c|\gamma\rangle \geqslant \langle \varphi \ominus \psi|\gamma\rangle = \langle \varphi|\mu\rangle - \langle \psi|\nu\rangle,$$

implying that $\varphi \ominus \psi = c$ γ a.e. $\square$

The proof of Theorem 5 also relies on a few elementary lemmas from measure theory.

Lemma 9. *If* μ_N *converges weakly to* μ*, then for any point* $x \in \mathrm{spt}(\mu)$ *there exists a sequence* $x_N \in \mathrm{spt}(\mu_N)$ *converging to* x*.*

Proof. Consider $x \in \mathrm{spt}(\mu)$. For any $k \in \mathbb{N}$, consider the function $\varphi_k(z) = \max(1 - kd(x, z), 0)$, in $\mathcal{C}^0(X)$. Then,

$$\lim_{N\to\infty} \langle \varphi_k|\mu_N\rangle = \langle \varphi_k|\mu\rangle > 0,$$

where the last inequality holds because x belongs to the support of μ. Then, there exists N_k such that for any $N \geqslant N_k$, $\langle \varphi_k|\mu_N\rangle > 0$, implying the existence of $x_N \in X$ such that $x_N \in \mathrm{spt}(\mu_N)$ and $d(x_N, x) \leqslant 1/k$. By a diagonal argument, this allows to construct a sequence of points $(x_N)_{N\in\mathbb{N}}$ such that $x_N \in \mathrm{spt}(\mu_N)$ and $\lim_{N\to+\infty} x_N = x$. $\square$

Lemma 10. *Let* X *be a compact space and* $\mu \in \mathcal{P}(X)$*. Then, there exists a sequence of finitely supported probability measures weakly converging to* μ*.*

Proof. For any $\varepsilon > 0$, by compactness there exists N points $x_1, \dots, x_N$ such that $X \subseteq \bigcup_i \mathrm{B}(x_i, \varepsilon)$. We define a partition $K_1, \dots, K_N$ of X recursively by $K_i = \mathrm{B}(x_i, \varepsilon) \setminus (K_1 \cup ... \cup K_{i-1})$ and we introduce

$$\mu_\varepsilon := \sum_{1\leqslant i\leqslant N} \mu(K_i)\delta_{x_i}.$$

To prove weak convergence of μ_ε to μ as $\varepsilon \to 0$, take $\varphi \in \mathcal{C}^0(X)$. By compactness of X, φ admits a modulus of continuity ω, i.e. $\lim_{t\to0} \omega(t) = 0$ and $|\varphi(x) - \varphi(y)| \leqslant \omega(d(x, y))$. Using that $\mathrm{diam}(K_i) \leqslant \varepsilon$, we get

$$\left|\int \varphi \mathrm{d}\mu - \int \varphi \mathrm{d}\mu_\varepsilon\right| = \left|\sum_{1\leqslant i\leqslant N} \int_{K_i} \varphi(x) - \varphi(x_i)\mathrm{d}\mu\right| \leqslant \omega(\varepsilon),$$

We deduce $\lim_{\varepsilon\to0}\langle \varphi|\mu_\varepsilon\rangle = \langle \varphi|\mu\rangle$, so that μ_ε weakly converges to μ. $\square$

Lemma 11. *If* $(\mu, \nu) \in \mathcal{P}(X) \times \mathcal{P}(Y)$ *are finitely supported,* (KP) = (DP)*.*

Proof. Assume that $\mu = \sum_{1\leqslant i\leqslant N} \mu_i\delta_{x_i}, \nu = \sum_{1\leqslant j\leqslant M} \nu_j\delta_{y_j}$, where all the μ_i and μ_j are strictly positive, and consider the linear programming problem

$$(\mathrm{KP})' = \min \left\{ \sum_{i,j} \gamma_{ij} c(x_i, y_j) \mid \gamma_{ij} \geqslant 0, \sum_j \gamma_{ij} = \mu_i, \sum_i \gamma_{ij} = \nu_j \right\},$$

which admits a solution which we denote γ. By Karush-Kuhn-Tucker theorem, there exists Lagrange multipliers $(\varphi_i)_{1\leqslant i\leqslant N}$, $(\psi_j)_{1\leqslant j\leqslant M}$ and $(\pi_{ij})_{1\leqslant i\leqslant N, 1\leqslant j\leqslant M}$ such that

$$\begin{cases} \varphi_i - \psi_j - c(x_i, y_j) = \pi_{ij} \\ \gamma_{ij}\pi_{ij} = 0 \\ \pi_{ij} \leqslant 0 \end{cases}$$

In particular, $\varphi_i - \psi_j \leqslant c(x_i, y_j)$ with equality if $\gamma_{ij} > 0$. To prove strong duality between the original problems (KP) and (DP), we construct two functions $\hat\varphi, \hat\psi$ such that $\hat\varphi \ominus \hat\psi \leqslant c$ with equality on the set $\{(x_i, y_j) \mid \gamma_{ij} > 0\}$. For this purpose, we first introduce

$$\psi(y) = \begin{cases} \psi_i & \text{if } y = y_i \\ +\infty & \text{if not} \end{cases}$$

and let $\hat\varphi = \psi^c$, $\hat\psi = \hat\varphi^c$. Let $i \in \{1, \cdots, M\}$. Since $\mu_i = \sum_j \gamma_{ij} \neq 0$, there exists $j \in \{1, \ldots, M\}$ such that $\gamma_{ij} > 0$. Using $\gamma_{ij}\pi_{ij} = 0$, we deduce that so that $\varphi_i - \psi_j = c(x_i, y_j)$, giving

$$\hat\varphi(x_i) = \min_{k\in\{1,\ldots,N\}} c(x_i, y_k) + \psi_k = c(x_i, y_j) + \psi_j = \varphi_i.$$

Similarly, one can show that $\hat\psi(y_j) = \psi_j$ for all $j \in \{1, \ldots, M\}$. Finally, define $\gamma = \sum_{ij} \gamma_{ij}\delta_{(x_i,y_j)} \in \Gamma(\mu, \nu)$. Then one can check that $\hat\varphi \ominus \hat\psi \leqslant c$ with equality γ-a.e., so that (KP) = (DP) by Proposition 8. $\square$

Proof of Theorem 5. By Lemma 10, there exists a sequence $\mu_k \in \mathcal{P}(X)$ (resp. $\nu_k \in \mathcal{P}(Y)$) of finitely supported measures which converge weakly to μ (resp. ν). We denote $(\mathrm{KP})_k$ and $(\mathrm{DP})_k$ the primal and dual Kantorovich problems between μ_k and ν_k. By Proposition 3, there exists a solution (φ_k, ψ_k) of $(\mathrm{DP})_k$, such that $\varphi_k = \psi_k^c$ and $\psi_k = \varphi_k^c$. Moreover, since strong duality holds for finitely supported measures (Lemma 11), we see (Proposition 8) that γ_k is supported on the set

$$S_k = \{(x, y) \in X \times Y \mid \varphi_k(x) - \psi_k(y) = c(x, y)\}.$$

Adding a constant if necessary, we can also assume that $\varphi_k(x_0) = 0$ for some point $x_0 \in X$. As c-concave functions, φ_k and ψ_k have the same modulus of continuity as the cost function c (see Lemma 4), and they are uniformly bounded (using $\varphi_k(x_0) = 0$). Using Arzelà-Ascoli theorem, we can therefore assume that up to subsequences, (φ_k) (resp. (ψ_k)) converges to some φ (resp ψ) uniformly. Then, one easily sees that $\varphi \ominus \psi \leqslant c$ so that (φ, ψ) are admissible for the dual problem (DP).

By compactness of $\mathcal{P}(X \times Y)$, we can assume that the sequence $\gamma_k \in \Gamma(\mu_k, \nu_k)$ converges to some $\gamma \in \Gamma(\mu, \nu)$. Moreover, by Lemma 9, every pair $(x, y) \in \mathrm{spt}(\gamma)$ can be approximated by a sequence of pairs $(x_k, y_k) \in \mathrm{spt}(\gamma_k)$ i.e. $\lim_{k\to\infty}(x_k, y_k) = (x, y)$. Since γ_k is supported on S_k one has $c(x_k, y_k) =$

$\varphi_k(x_k) - \psi_k(x_k)$, which gives at the limit $c(x,y) = \varphi(x) - \psi(y)$. We have just shown that for every point pair (x,y) in $\mathrm{spt}(\gamma)$, $c(x,y) = \varphi(x) - \psi(y)$ where φ, ψ is admissible. By Proposition 8, this shows that γ and (φ, ψ) are optimal for their respective problems and that (KP) = (DP). □

Corollary 6 is a direct consequence of Proposition 8 and of the strong duality (KP) = (DP).

Solution of Monge's problem for Twisted costs. We now show how to use Kantorovich duality to prove the existence of optimal transport maps when the source measure is absolutely continuous on a compact subset of $\mathbb{R}^d$ and when the cost function satisfies the following condition:

Definition 8 (Twisted cost)**.** Let $\Omega_X, \Omega_Y \subseteq \mathbb{R}^d$ be open subsets, and $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$. The cost function satisfies the twist condition if

$$\forall x_0 \in \Omega_X, \text{ the map } y \in \Omega_Y \mapsto v := \nabla_x c(x_0, y) \in \mathbb{R}^d \text{ is injective,} \tag{2.23}$$

where $\nabla_x c(x_0, y)$ denotes the gradient of $x \mapsto c(\cdot, y)$ at $x = x_0$. Given $x_0 \in \Omega_X$ and $v \in \mathbb{R}^d$, we denote $y_c(x_0, v)$ the unique point (if it exists) such that $\nabla_x c(x_0, y_c(x_0, v)) = v$. The map $v \mapsto y_c(x_0, v)$ is often called the c-exponential map at x_0.

Example 4 (Quadratic cost). Let $c(x,y) = \|x-y\|^2$. Then, for any $x_0 \in X$, the map $y \mapsto \nabla_x c(x_0, y) = 2(x_0 - y)$ is injective, so that c satisfies the twist condition. Moreover, given $v \in \mathbb{R}^d$, the unique y such that $\nabla_x c(x_0, y) = 2(x_0 - y) = v$ is $y = x_0 - \frac{1}{2}v$, implying that $y_c(x_0, v) = x_0 - \frac{1}{2}v$.

The following theorem is due to Brenier [19] in the case of the quadratic cost (i.e. $c(x,y) = \|x-y\|^2$) and Gangbo-McCann in the general case of twisted costs [51].

Given $X \subseteq \Omega_X \subset \mathbb{R}^d$, we define $\mathcal{P}^{\mathrm{ac}}(X)$ as the set of probability measures on Ω_X that are absolutely continuous with respect to the Lebesgue measure, and with support included in X.

Theorem 12 (Brenier [19], Gangbo-McCann [51])**.** *Let $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$ be a twisted cost, let $X \subseteq \Omega_X, Y \subseteq \Omega_Y$ be compact sets, and let $(\mu, \nu) \in \mathcal{P}^{\mathrm{ac}}(X) \times \mathcal{P}(Y)$. Then, there exists a c-concave function $\varphi \in \mathrm{Lip}(X)$ such that $\nu = T_\# \mu$ where $T(x) = y_c(x, \nabla\varphi(x))$. Moreover, the only optimal transport plan between μ and ν is γ_T.*

Example 5. If $h \in \mathcal{C}^1(\mathbb{R}^d)$ is strictly convex, in particular if $h(x) = \|x\|^p$, then the map $x \mapsto \nabla h(x)$ is injective. Take $c(x,y) = h(x-y)$, so that $y \mapsto \nabla_x c(x,y) = \nabla_x h(x-y) = \nabla h(x-y)$ is also injective. Moreover, given $x_0 \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$, the unique solution y to $v = \nabla h(x_0 - y)$ is $y = y_c(x_0, v) := x_0 - (\nabla h)^{-1}(v)$. As a consequence, under the hypothesis of the theorem above, the transport map is of the form

$$T(x) = x - (\nabla h)^{-1}(\nabla \varphi(x))$$

where φ is a c-convex function.

The following lemma shows that a transport plan is induced by a transport map if it is concentrated on the graph of a map.

Lemma 13. *Let $\gamma \in \Gamma(\mu, \nu)$ and $T : X \to Y$ measurable be such that $\gamma(\{(x, y) \in X \times Y \mid T(x) \neq y\}) = 0$. Then, $\gamma = \gamma_T$.*

Proof. By definition of γ_T one has $\gamma_T(A \times B) = \mu(T^{-1}(B) \cap A)$ for all Borel sets $A \subseteq X$ and $B \subseteq Y$. On the other hand,

$$\begin{aligned}\gamma(A \times B) &= \gamma(\{(x, y) \mid x \in A, \text{ and } y \in B\}) \\ &= \gamma(\{(x, y) \mid x \in A, y \in B \text{ and } y = T(x)\}) \\ &= \gamma(\{(x, y) \mid x \in A \cap T^{-1}(B), y = T(x)\} \\ &= \mu(A \cap T^{-1}(B)),\end{aligned}$$

thus proving the claim. □

Proof of Theorem 12. Enlarging X if necessary (while keeping it compact and inside Ω_X), we may assume that $\mathrm{spt}(\mu)$ is contained in the interior of X. First note that by compactness of $X \times Y$ and since c is $\mathcal{C}^1$, the cost c is Lipschitz on $X \times Y$. Take $(\varphi, \varphi^{\overline{c}})$ a maximizing pair for (DP) with φ c-concave. By the formula $\varphi(x) = \min_{y \in Y} c(x, y) + \varphi^{\overline{c}}(y)$ one can see that φ is Lipschitz. By Rademacher theorem, φ is differentiable Lebesgue almost everywhere, and by the hypothesis $\mu \in \mathcal{P}^{\mathrm{ac}}(X)$, it is therefore differentiable on a set $B \subseteq \mathrm{spt}(\mu)$ with $\mu(B) = 1$. Consider an optimal transport plan $\gamma \in \Gamma(\mu, \nu)$. For every pair of points $(x_0, y_0) \in \mathrm{spt}(\gamma) \cap B \times Y$, we have

$$\forall x \in X, \varphi(x) - c(x, y_0) \leqslant \varphi^{\overline{c}}(y_0)$$

with equality at $x = x_0$, so that x_0 maximizes the function $\varphi - c(\cdot, y_0)$. Since $x_0 \in \mathrm{spt}(\mu)$, x_0 belongs to the interior of X, one necessarily has $\nabla\varphi(x_0) = \nabla_x c(x_0, y_0)$. Then, by the twist condition, one necessarily has $y_0 = y_c(x_0, \nabla\varphi(x_0))$. This shows that any optimal transport plan γ is supported on the graph of the map $T : x \in B \mapsto y_c(x_0, \nabla\varphi(x_0))$, and $\gamma = \gamma_T$ by the previous lemma. □

We finish this section with a stability result for optimal transport maps (a more general result can be found in [98, Chapter 5]).

Proposition 14 (Stability of OT maps)**.** *Let $X \subseteq \Omega_X$ and $Y \subseteq \Omega_Y$ be compact subsets of open sets $\Omega_X, \Omega_Y \subseteq \mathbb{R}^d$, and $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$ be a twisted cost. Let $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$, and let $(\mu_k) \in \mathcal{P}(Y)$ be a sequence of measures converging weakly to $\mu \in \mathcal{P}(Y)$. Define T_k (resp. T) as the unique optimal transport map between ρ and μ_k (resp. ρ and μ). Then, $\lim_{k \to +\infty} \|T_k - T\|_{\mathrm{L}^1(\rho)} = 0$.*

Remark 4. Note that unlike the stability theorem for optimal transport plans (Theorem 7), the convergence in Proposition 14 is for the whole sequence and not up to subsequence. This theorem is not quantitative, and there exists very few quantitative variants of this theorem. We are aware of two such results. To state them, given a fixed $\rho \in \mathcal{P}(X)$ and $\mu \in \mathcal{P}(Y)$, we denote T_μ the unique optimal transport map between ρ and μ.

- A first result of Ambrosio, reported in an article of Gigli [53, Proposition 3.3 and Corollary 3.4], shows that if μ_0 is such that the optimal transport map T_{μ_0} is Lipschitz, then

$$\|T_\mu - T_{\mu_0}\|^2_{\mathrm{L}^2(\rho)} \leqslant C(\mu_0)\,\mathrm{W}_2(\mu, \mu_0).$$

The theorem in [53] holds for the quadratic cost on $\mathbb{R}^d$. It was recently generalized to other cost functions [5].

- Berman [15] proves a global estimate, not assuming the regularity of T_{μ_0} but with a worse Hölder exponent, of the form
$$\|T_\mu - T_{\mu_0}\|^2_{\mathrm{L}^2(\rho)} \leqslant C\,\mathrm{W}_1(\mu,\mu_0)^{1/2^{d-1}},$$
assuming that ρ is bounded from below on a compact convex domain of $\mathbb{R}^d$, when the cost is quadratic. The constant then C only depends on X, Y and ρ. Recently a similar bound with an exponent independent on the dimension was obtained by Mérigot, Delalande and Chazal [71]:
$$\|T_\mu - T_{\mu_0}\|^2_{\mathrm{L}^2(\rho)} \leqslant C\,\mathrm{W}_1(\mu,\mu_0)^{1/15}.$$

Proof. As before, without loss of generality, we assume that $\mathrm{spt}(\sigma)$ lies in the interior of X. Let (φ_k, ψ_k) be solutions to $(\mathrm{DP})_k$, which are c-conjugate to each other, and such that $\varphi_k(x_0) = 0$ for some $x_0 \in X$. Then, by stability of Kantorovich potentials, there exists a subsequence (φ_k, ψ_k) (which we do not relabel) which converges uniformly to (φ, ψ). Moreover, (φ, ψ) are Kantorovich potentials for (DP), and are also c-conjugate to each other.

Since $\varphi, \varphi_k \in \mathrm{Lip}(X)$ are differentiable almost everywhere, there exists a subset $Z \subseteq \mathrm{spt}(\sigma)$ with $\mu(Z) = 1$ and such that for all $x \in Z$, $\nabla\varphi_k$ exists for all k and $\nabla\varphi$ exists. Let $x \in Z$. Using
$$\varphi_k(x) - \psi_k(T_k(x)) = c(x, T_k(x)),$$
we get that for any cluster point y of the sequence $(T_k(x))_k$,
$$\begin{cases} \varphi(x) - \psi(y) = c(x,y), \\ \varphi(x') - \psi(y) \leqslant c(x',y) \quad \forall x' \in X \end{cases},$$
where the second inequality is obtained using $\varphi_k \ominus \psi_k \leqslant c$. Thus, as in the proof of Brenier-McCann-Gangbo's theorem, x is a minimizer of $c(\cdot, y) - \varphi$, i.e. $\nabla_x c(x,y) = \nabla\varphi(x)$, implying that $y = y_c(x, \nabla\varphi(x)) = T(x)$. By compactness, this shows that the whole sequence $(T_k(x))_k$ converges to $S(x)$. Therefore, T_k converges σ-almost everywhere to T, and $\mathrm{L}^1(\sigma)$ convergence follows easily. □

2.3. **Kantorovich's functional.** As already mentioned in Equation (2.20), the Kantorovich's dual problem (DP) can be expressed as an unconstrained maximization problem:
$$(\mathrm{DP}) = \max_{\psi \in \mathcal{C}^0(Y)} \int_X \psi^c \mathrm{d}\mu - \int_Y \psi \mathrm{d}\nu.$$
This motivates the definition of *Kantorovich's functional* as follows

Definition 9. The Kantorovitch functional is defined on $\mathcal{C}^0(Y)$ by
$$\mathcal{K}(\psi) = \int_X \psi^c \mathrm{d}\mu - \int_Y \psi \mathrm{d}\nu. \tag{2.24}$$

The Kantorovitch dual problem therefore amounts to maximizing the Kantorovitch functional:
$$(\mathrm{DP}) = \max_{\psi \in \mathcal{C}^0(Y)} \mathcal{K}(\psi).$$

This subsection is devoted to the general computation of the superdifferential of Kantorovich's functional when Y is finite. This computation will be used to construct and study algorithms for discretized optimal transport problems. The definition, as well as basic properties on the superdifferential $\partial^+ F$ of a function F are recalled in Appendix 5.1.

Proposition 15. *Let X be a compact space, Y be finite, $c \in \mathcal{C}^0(X \times Y)$ and $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Then, for all $\psi_0 \in \mathbb{R}^Y$,*

$$\partial^+ \mathcal{K}(\psi_0) = \{\Pi_{Y\#}\gamma - \nu \mid \gamma \in \Gamma_{\psi_0}(\mu)\}. \tag{2.25}$$

where $\Gamma_{\psi_0}(\mu)$ is the set of probability measures on $X \times Y$ with first marginal μ and supported on the c-subdifferential $\partial^c\psi_0$ (defined in Eq. (2.21)), i.e.

$$\Gamma_{\psi_0}(\mu) = \{\gamma \in \mathcal{P}(X \times Y) \mid \Pi_{X\#}\gamma = \mu \text{ and } \mathrm{spt}(\gamma) \subseteq \partial^c\psi_0\}. \tag{2.26}$$

Proof. Let $\gamma \in \Gamma_{\psi_0}(\mu)$. Then, for all $\psi \in \mathbb{R}^Y$,

$$\begin{aligned}
\mathcal{K}(\psi) &= \int \psi^c \mathrm{d}\mu - \int \psi \mathrm{d}\nu \\
&= \int \psi^c(x) \mathrm{d}\gamma(x,y) - \int \psi \mathrm{d}\nu \\
&\leqslant \int c(x,y) + \psi(y) \mathrm{d}\gamma(x,y) - \int \psi \mathrm{d}\nu,
\end{aligned}$$

where we used $\Pi_{X\#}\gamma = \mu$ to get the second equality and $\psi^c(x) \leqslant c(x,y) + \psi(y)$ to get the inequality. Note also that equality holds if $\psi = \psi_0$, by assumption on the support of γ. Hence,

$$\begin{aligned}
\mathcal{K}(\psi) &\leqslant \mathcal{K}(\psi_0) + \int (\psi(y) - \psi_0(y)) \mathrm{d}\gamma(x,y) - \int (\psi - \psi_0) \mathrm{d}\nu \\
&= \mathcal{K}(\psi_0) + \langle \Pi_{Y\#}\gamma - \nu | \psi - \psi_0 \rangle.
\end{aligned}$$

This implies by definition that $\Pi_{Y\#}\gamma - \nu$ lies in the superdifferential $\partial^+\mathcal{K}(\psi_0)$, giving us the inclusion

$$D(\psi_0) := \{\Pi_{Y\#}\gamma - \nu \mid \gamma \in \Gamma_{\psi_0}(\mu)\} \subseteq \partial^+\mathcal{K}(\psi_0).$$

Note also that the superdifferential of $\mathcal{K}$ is non-empty at any $\psi_0 \in \mathbb{R}^Y$, so that $\mathcal{K}$ is concave. As a concave function on the finite-dimensional space $\mathbb{R}^Y$, $\mathcal{K}$ is differentiable almost everywhere and one has $\partial\mathcal{K}^+(\psi) = \{\nabla\mathcal{K}(\psi)\}$ at differentiability points.

We now show that $\partial\mathcal{K}^+(\psi_0) \subset D(\psi_0)$, using the characterization of the subdifferential recalled in the Appendix:

$$\partial\mathcal{K}^+(\psi_0) = \mathrm{conv}\left\{\lim_{n\to\infty} \nabla\mathcal{K}(\psi^n) \mid (\psi^n)_{n\in\mathbb{N}} \in S\right\},$$

where S is the set of sequences $(\psi^n)_{n\in\mathbb{N}}$ that converge to ψ_0, such that $\nabla\mathcal{K}(\psi^n)$ exist and admit a limit as $n \to +\infty$. Let $v = \lim_{n\to\infty} \nabla\mathcal{K}(\psi^n)$, where $(\psi^n)_{n\in\mathbb{N}}$ belongs to the set S. For every n, there exists $\gamma^n \in \Gamma_{\psi^n}(\mu)$ such that $\nabla\mathcal{K}(\psi^n) = v^n := \Pi_{Y\#}\gamma^n - \nu$. By compactness of $\mathcal{P}(X \times Y)$, one can assume (taking a subsequence if necessary) that γ^n weakly converges to some γ, and it is not difficult to check that $\gamma \in \Gamma_{\psi_0}(\mu)$, ensuring that the sequence v^n converges to some $v \in D(\psi_0)$. Thus,

$$\left\{\lim_{n\to\infty} \nabla\mathcal{K}(\psi^n) \mid (\psi^n)_{n\in\mathbb{N}} \in S\right\} \subseteq D(\psi_0).$$

Taking the convex hull and using the convexity of $D(\psi_0)$, we get $\partial^+\mathcal{K}(\psi_0) \subseteq D(\psi_0)$ as desired. $\square$

As a corollary of this proposition, we obtain an explicit expression for the left and right partial deriatives of $\mathcal{K}$, and a characterization of its differentiability. In this corollary, we use the terminology of semi-discrete optimal transport (Section 4.1), and we will refer to the c-subdifferential at $y \in Y$ as *Laguerre cell associated to* y and we will denote it by $\mathrm{Lag}_y(\psi)$.

$$\mathrm{Lag}_y(\psi) := \{x \in X \mid \forall z \in Y, c(x,y) + \psi(y) \leqslant c(x,z) + \psi(z)\}. \tag{2.27}$$

We also need to introduce the *strict Laguerre cell* $\mathrm{SLag}_y(\psi)$:

$$\mathrm{SLag}_y(\psi) := \{x \in X \mid \forall z \in Y, c(x,y) + \psi(y) < c(x,z) + \psi(z)\}. \tag{2.28}$$

Corollary 16 (Directional derivatives of $\mathcal{K}$)**.** *Let $\psi \in \mathbb{R}^Y$, $y \in Y$ and define $\kappa(t) = \mathcal{K}(\psi^t)$ where $\psi^t = \psi + t\mathbf{1}_y$. Then, κ is concave and*

$$\partial^+\kappa(t) = [\mu(\mathrm{SLag}_y(\psi^t)) - \nu(\{y\}), \mu(\mathrm{Lag}_y(\psi^t)) - \nu(\{y\})]$$

In particular $\mathcal{K}$ is differentiable at $\psi \in \mathbb{R}^Y$ iff $\mu(\mathrm{Lag}_y(\psi) \setminus \mathrm{SLag}_y(\psi)) = 0$ for all $y \in Y$, and in this case

$$\nabla\mathcal{K}(\psi) = \Big(\mu(\mathrm{Lag}_y(\psi)) - \nu(\{y\})\Big)_{y\in Y}.$$

Proof. Using Hahn-Banach's extension theorem, one can easily see that the super-differential of κ at t is the projection of the super-differential $\mathcal{K}$ at ψ^t:

$$\partial^+\kappa(t) = \left\{\langle \pi | \mathbf{1}_y\rangle \mid \pi \in \partial^+\mathcal{K}(\psi^t)\right\}.$$

Combining with the previous proposition we get

$$\begin{aligned}\partial^+\kappa(t) &= \left\{\langle \Pi_{X\#\gamma} - \nu | \mathbf{1}_y\rangle \mid \gamma \in \Gamma_{\psi^t}(\mu)\right\}\\ &= \left\{\gamma(X \times \{y\}) - \nu(\{y\}) \mid \gamma \in \Gamma_{\psi^t}(\mu)\right\}\end{aligned}$$

To obtain the desired formula for $\partial\kappa^+(t)$, it remains to prove that

$$\max\left\{\gamma(X \times \{y\}) \mid \gamma \in \Gamma_{\psi^t}(\mu)\right\} = \mu(\mathrm{Lag}_y(\psi^t)),$$
$$\min\left\{\gamma(X \times \{y\}) \mid \gamma \in \Gamma_{\psi^t}(\mu)\right\} = \mu(\mathrm{SLag}_y(\psi^t)).$$

We only prove the first equality, the second one being similar. Denote $Z = X \setminus \mathrm{Lag}_y(\psi^t)$, so that for any $\gamma \in \Gamma_{\psi^t}(\mu)$,

$$\gamma(X \times \{y\}) = \gamma(\mathrm{Lag}_y(\psi^t) \times \{y\}) + \gamma(Z \times \{y\}).$$

Moreover, by definition of $\mathrm{Lag}_y(\psi^t)$, $Z \times \{y\} \cap \partial^c\psi^t = \emptyset$. Since γ belongs to $\Gamma_{\psi^t}(\mu)$, we have $\mathrm{spt}(\gamma) \subseteq \partial^c\psi^t$ so that $\gamma(Z \times \{y\}) = 0$. This gives us

$$\gamma(X \times \{y\}) = \gamma(\mathrm{Lag}_y(\psi^t) \times \{y\}) \leqslant \gamma(\mathrm{Lag}_y(\psi^t) \times Y) = \mu(\mathrm{Lag}_y(\psi^t)),$$

where we used $\Pi_{X\#}\gamma = \mu$ to get the last equality. This proves that

$$\sup\left\{\gamma(X \times \{y\}) \mid \gamma \in \Gamma_{\psi^t}(\mu)\right\} \leqslant \mu(\mathrm{Lag}_y(\psi^t)).$$

To show equality, we consider an explicit $\gamma \in \Gamma_{\psi^t}(\mu)$ using a map $T : X \to Y$ defined as follows: for $x \in \mathrm{Lag}_y(\psi^t)$, we set $T(x) = y$ and for points $x \notin \mathrm{Lag}_y(\psi^t)$, we define $T(x)$ to be an arbitrary $z \in Y$ such that $x \in \mathrm{Lag}_z(\psi^t)$. Then, one can readily check that $\gamma = (\mathrm{id}, T)_{\#}\mu$ belongs to $\Gamma_{\psi^t}(\mu)$ and that by construction, $\gamma(X \times \{y\}) = \mu(\mathrm{Lag}_y(\psi^t))$. $\square$

3. Discrete optimal transport

In this part we present two algorithms for solving discrete optimal transport problems, wich can be both be interpreted using Kantorovich's duality:

- The first one is Bertsekas' auction algorithm, which allows to solve optimal transport problem where the source and targed measures are uniform over two sets with the same cardinality, a case known as the *assignment problem* in combinatorial optimization. Bertsekas' algorithm is a *coordinate-ascent method* that iteratively modifies the coordinates of the dual variable ψ so as to reach a maximizer of the Kantorovitch functional $\mathcal{K}(\psi)$.
- The second algorithm is the *Sinkhorn-Knopp's algorithm* that allows to solve the entropic regularization of (discrete) optimal transport problems. This algorithm can be seen as a *block-coordinate ascent method* since it amounts to maximizing the dual of the regularized optimal transport problem, denoted by $\mathcal{K}^\eta(\varphi, \psi)$, by alternatively optimizing with respect to the two dual variables φ and ψ.

3.1. **Formulation of discrete optimal transport.**

Primal and dual problems. We consider in this section that the two sets X and Y are finite, and we consider two discrete probability measures $\mu = \sum_{x\in X} \mu_x \delta_x$ and $\nu = \sum_{y \in Y} \nu_y \delta_y$. This setting occurs frequently in applications. The set of transport plans is then given by

$$\Gamma(\mu, \nu) = \left\{ \gamma = \sum_{x,y} \gamma_{x,y} \delta_{(x,y)} \mid \gamma_{x,y} \geqslant 0, \sum_{y\in Y} \gamma_{x,y} = \mu_x, \sum_{x \in X} \gamma_{x,y} = \nu_y \right\},$$

and is often referred to as the *transportation polytope*. In this discrete setting, we will conflate a transport plan $\gamma \in \Gamma(\mu,\nu)$ with the matrix $(\gamma_{x,y})_{(x,y)\in X\times Y}$, which formally is the density of γ with respect to the counting measure. The constraint $\sum_y \gamma_{x,y} = \mu_x$ encodes the fact that all mass from x is transported somewhere in Y, while the constraint $\sum_x \gamma_{x,y} = \nu_y$ tells us that the mass at y is transported from somewhere in X. The Kantorovitch problem for a cost function $c : X \times Y \to \mathbb{R}$ then reads

$$(\mathrm{KP}) = \min_{\gamma \in \Gamma(\mu,\nu)} \sum_{x\in X, y \in Y} c(x,y)\gamma_{x,y}. \tag{3.29}$$

As seen in Section 2.2, the dual (DP) of this linear programming problem amounts to maximizing the Kantorovitch functional $\mathcal{K}$ (2.24), which in this setting can be expressed as

$$\mathcal{K}(\psi) = \sum_{x\in X} \min_{y\in Y} (c(x,y) + \psi(y))\mu_x - \sum_{y \in Y} \psi(y)\nu_y, \tag{3.30}$$

where $\psi \in \mathbb{R}^Y$ is a function over the finite set Y. Since strong duality holds (Theorem 5), one has

$$(\mathrm{KP}) = (\mathrm{DP}) = \max_{\psi\in\mathbb{R}^Y} \mathcal{K}(\psi).$$

Remark 5. Knowing a maximizer $\psi \in \mathbb{R}^Y$ of $\mathcal{K}$ does not directly allow to recover an optimal transport plan γ. However, by Corollary 6, we know that any transport plan $\gamma \in \Gamma(\mu, \nu)$ is optimal if and only if its support is included in the c-subdifferential of ψ:

$$\mathrm{spt}(\gamma) \subset \partial^c \psi = \Big\{(x, y) \in X \times Y \mid \psi^c(x) = c(x, y) + \psi(y)\Big\}.$$

3.2. **Linear assignment via coordinate ascent.**

Assignment problem. When the two sets X and Y have the same cardinal N and when μ and ν are uniform probability measures over these sets, namely

$$\mu = \frac{1}{N} \sum_{x \in X} \delta_x, \quad \nu = \frac{1}{N} \sum_{y \in Y} \delta_y, \tag{3.31}$$

then Monge's problem corresponds to the *(linear) assignment problem* (AP) which is one of the most famous combinatorial optimization problem:

$$(\mathrm{AP}) = \min \left\{ \frac{1}{N} \sum_{x \in X} c(x, \sigma(x)) \mid \sigma : X \to Y \text{ is a bijection } \right\}. \tag{3.32}$$

This problem and its variants have generated a very important amount of research, as demonstrated by the bibliography of the book by Burkard, Dell'Amico and Martello on this topic [21].

Note that the set of bijections from X to Y has cardinal $N!$, making it practically impossible to solve (AP) through direct enumeration. Using Birkhoff's theorem on bistochastic matrices, we will show that the assignment problem coincides with the Kantorovitch problem.

Definition 10 (Bistochastic matrices)**.** A N-by-N bistochastic matrix is a square matrix $M \in \mathcal{M}_N(\mathbb{R})$ with non-negative coefficients such that the sum of any row and any column equals one:

$$\forall i \in \{1, \dots, N\}, \sum_j M_{ij} = 1, \quad \forall j \in \{1, \dots, N\}, \sum_i M_{ij} = 1$$

We denote the set of N-by-N bistochatic matrices as $\mathcal{B}_N \subseteq \mathcal{M}_N(\mathbb{R})$.

Definition 11 (Permutation matrix)**.** The set of permutations (bijections) from $\{1, \dots, N\}$ to itself is denoted $\mathfrak{S}_N$. Given a permutation $\sigma \in \mathfrak{S}_N$, we associate the permutation matrix

$$M[\sigma]_{ij} = \begin{cases} 1 \text{ if } \sigma(i) = j \\ 0 \text{ if not} \end{cases}.$$

One can easily check that if σ is a permutation, then $M[\sigma]$ belongs to $\mathcal{B}_N$. Birkhoff's theorem on the other hand asserts that the extremal points of the polyhedron $\mathcal{B}_N$ are permutation matrices, implying thanks to Krein-Milman theorem that every bistochastic matrix can be obtained as a (finite) convex combination of permutation matrices.

Theorem 17 (Birkhoff)**.** *The extremal points of $\mathcal{B}_N$ are the permutation matrices. In particular, $\mathcal{B}_N = \mathrm{conv}\{M[\sigma] \mid \sigma \in \mathfrak{S}_N\}$.*

Kantorovitch's problem (KP) amounts to minimizing a linear function over the set of bistochastic matrices which is convex. Birkoff's theorem implies that there exists a bijection that solves this problem (KP), hence the following theorem:

Theorem 18. *Let μ and ν be as in* (3.31). *Then,* (AP) = (KP).

Proof. Take an arbitrary ordering of the points in X and Y, i.e. $X = \{x_1, \dots, x_N\}$ and $Y = \{y_1, \dots, y_N\}$. Then $\gamma \in \mathbb{R}^{X\times Y} \simeq \mathcal{M}_N(\mathbb{R})$ is a transport plan between μ and ν iff $N\gamma \in \mathcal{B}_N$. Since bistochastic matrices include permutation matrices, we have (KP) $\leqslant$ (AP), and the converse follows from the fact that the minimum in (KP) is attained at an extreme point of $\mathcal{B}_N$, i.e. a permutation matrix. □

Dual coordinate ascent methods. We follow Bertsekas [16] by trying to solve the assignment problem (AP) through the unconstrained dual problem (3.1). Combining Theorem 18 and Corollary 6, we have the following proposition.

Proposition 19. *The following statements are equivalent*

- *ψ is a global maximizer of the Kantorovitch functional $\mathcal{K}$*
- *There exists a bijection $\sigma : X \to Y$ that satisfies*

$$\forall x \in X,\ c(x, \sigma(x)) + \psi(\sigma(x)) = \min_{y\in Y} c(x, y) + \psi(y).$$

A bijection σ satisfying this last equation is a solution to the linear assignment problem.

The idea of Bertsekas [16] is to iteratively modify the weights $\psi \in \mathbb{R}^Y$ so as to reach a maximizer of $\mathcal{K}$. By Corollary 16, the gradient of the Kantorovitch functional, when it exists, is given by

$$\nabla\mathcal{K}(\psi) = \frac{1}{N}\Big(\mathrm{card}(\mathrm{Lag}_y(\psi)) - 1\Big)_{y\in Y}.$$

In addition, recalling the definition of a Laguerre cell,

$$\mathrm{Lag}_y(\psi) = \{x \in X \mid \forall y' \in Y,\ c(x, y) + \psi(y) \leqslant c(x, y') + \psi(y')\},$$

one can see that $\mathrm{card}(\mathrm{Lag}_y(\psi))$ is obviously decreasing when $\psi(y)$ increases. Therefore, in order to maximize the concave function $\mathcal{K}$, it is natural to increase the weight $\psi(y)$ of any Laguerre cells that satisfy $\mathrm{card}(\mathrm{Lag}_y(\psi)) > 1$. In the following lemma, we calculate the optimal increment, which is known as the *bid.*

Lemma 20 (Bidding increment)**.** *Let $\psi \in \mathbb{R}^Y$ and $y_0 \in Y$ be such that $\mathrm{Lag}_{y_0}(\psi) \neq \emptyset$. Then the maximum of the function $t \to \mathcal{K}(\psi + t\mathbf{1}_{y_0})$ is reached at*

$$\mathrm{bid}_{y_0}(\psi) = \max\{\mathrm{bid}_{y_0}(\psi, x),\ x \in \mathrm{Lag}_{y_0}(\psi)\},$$

where

$$\mathrm{bid}_{y_0}(\psi, x) := \left(\min_{y\in Y\setminus y_0} c(x, y) + \psi(y)\right) - (c(x, y_0) + \psi(y_0)).$$

Proof. Denote $\psi^t = \psi + t\mathbf{1}_{y_0}$. For $t > 0$ one has $\mathrm{Lag}_{y_0}(\psi^t) \subseteq \mathrm{Lag}_{y_0}(\psi)$. Remark also that for every $x \in X$, one has

$$\begin{aligned} x \in \mathrm{Lag}_{y_0}(\psi^t) &\Leftrightarrow \quad \forall z \neq y_0 \quad c(x, y_0) + \psi(y_0) + t \leqslant c(x, z) + \psi(z) \\ &\Leftrightarrow \quad t \leqslant \left(\min_{z \in Y \setminus y_0} c(x, z) + \psi(z)\right) - (c(x, y_0) + \psi(y_0)) \\ &\Leftrightarrow \quad t \leqslant \mathrm{bid}_{y_0}(\psi, x) \end{aligned}$$

This implies that $\mathrm{Lag}_{y_0}(\psi^t) \neq \emptyset$ if and only if $t \leqslant \mathrm{bid}_{y_0}(\psi)$. By Corollary 16, the upper-bound of the superdifferential $\partial^+\kappa(t)$ of the function $\kappa(t) = \mathcal{K}(\psi + t\mathbf{1}_{y_0})$ is $\mu(\mathrm{Lag}_{y_0}(\psi^t)) - \frac{1}{N}$. It is non-negative for $t \in [0, \mathrm{bid}_{y_0}(\psi)]$ and strictly negative for $t > \mathrm{bid}_{y_0}(\psi)$. This directly implies (for instance by (5.76)) that $0 \in \partial^+\kappa(\mathrm{bid}_{y_0}(\psi))$, so that the largest maximizer of κ is $\mathrm{bid}_{y_0}(\psi)$. □

Remark 6 (Economic interpretation of the bidding increment.). Assume that Y is a set of houses owned by one seller and X is a set of customers that want to buy a house. Given a set of prices $\psi : Y \to \mathbb{R}$, each customer $x \in X$ will make a compromise between the location of a house $y \in Y$ (measured by $c(x, y)$) and its price (measured by $\psi(y)$) by choosing a house among those minimizing $c(x, y) + \psi(y)$. In other words, x chooses y iff $x \in \mathrm{Lag}_y(\psi)$. Let y be a given house. The seller of y wants to maximize his profit, hence to increase $\psi(y)$ as much as possible while keeping (at least) one customer. Let $x \in X$ be a customer interested in the house y (i.e. $x \in \mathrm{Lag}_y(\psi)$). Then, $\mathrm{bid}_{y,x}(\psi)$ tells us how much it is possible to increase the price of y while keeping it interesting to x. The best choice for the seller is to increase the price by the maximum bid, which is the maximum raise so that there remains at least one customer, giving the definition of $\mathrm{bid}_y(\psi)$.

Remark 7 (Naive coordinate ascent). A naive algorithm would be to would choose at each step a coordinate $y \in Y$ such that $\mathrm{Lag}_y(\psi) \neq \emptyset$ and to increase $\psi(y)$ by the bidding increment $\mathrm{bid}_y(\psi)$. In practice, such an algorithm might get stuck at a point which is not a global maximizer, a phenomenon which is referred to as *jamming* in [17, §2]. In practice, this can happen when some bidding increments $\mathrm{bid}_y(\psi)$ vanishes, see Remark 8 below. Note that this is a particular case of the well known fact that coordinate ascent algorithms may converge to points that are not maximizers, when the maximized functional is nonsmooth.

In order to tackle the problem of non-convergence of coordinate ascent, Bertsekas and Eckstein changed the naive algorithm outlined above to impose that the bids are at least $\varepsilon > 0$. To analyse their algorithm, we introduce the notion of *ε-complementary slackness*, where ε can be seen as a tolerance.

Definition 12 (ε-Complementary slackness.)**.** A *partial assignment* is a couple (σ, S) where $S \subseteq X$ and $\sigma : S \to Y$ is an injective map. A partial assignment (σ, S) and a price function $\psi \in \mathbb{R}^Y$ satisfy *ε-complementary slackness* if for every x in S the following inequality holds:

$$c(x, \sigma(x)) + \psi(\sigma(x)) \leqslant \min_{y \in Y} [c(x, y) + \psi(y)] + \varepsilon. \tag{CS$_\varepsilon$}$$

In the economic interpretation, a partial assignment $\sigma : S \subseteq X \to Y$ satisfies (CS$_\varepsilon$) with respect to prices $\psi \in \mathbb{R}^Y$ if every customer $x \in S$ is assigned to a house $\sigma(x)$ which is "nearly optimal", i.e. is within ε of minimizing $c(x, \cdot) + \psi(\cdot)$ over Y.

Lemma 21. *If $\sigma : X \to Y$ is a bijection which satisfies* $(\mathrm{CS}_\varepsilon)$ *together with some $\psi \in \mathbb{R}^Y$, then*

$$(\mathrm{KP}) \leqslant \frac{1}{N} \sum_{x \in X} c(x, \sigma(x)) \leqslant (\mathrm{KP}) + \varepsilon. \tag{3.33}$$

Proof. The first inequality just comes from the fact σ is a particular transport plan. For the second inequality, by summing the $(\mathrm{CS}_\varepsilon)$ condition, one gets

$$\frac{1}{N} \sum_{x \in X} c(x, \sigma(x)) + \psi(\sigma(x)) \leqslant \frac{1}{N} \sum_{x \in X} \min_{y \in Y} (c(x, y) + \psi(y)) + \varepsilon.$$

This leads to

$$\frac{1}{N} \sum_{x \in X} c(x, \sigma(x)) \leqslant \mathcal{K}(\psi) + \varepsilon \leqslant (\mathrm{DP}) + \varepsilon = (\mathrm{KP}) + \varepsilon. \qquad \square$$

Bertsekas' auction algorithm. Bertsekas' auction algorithm maintains a partial matching (σ, S) and prices $\psi \in \mathbb{R}^Y$ that together satisfy ε-complementary slackness. At the end of the execution, σ is a bijection, and (σ, ψ) satisfy the ε-CS condition.

Algorithm 1 Bertsekas' auction algorithm

```
function AUCTION(c, ε, ψ = 0)
    S ← ∅                                                    ▷ All points are unassigned
    while ∃x ∈ X \ S do
        y0 ← arg min_{y∈Y} c(x, y) + ψ(y)
        y1 ← arg min_{y∈Y\{y0}} c(x, y) + ψ(y)
        ψ(y0) ← ψ(y0) + (c(x, y1) + ψ(y1)) − (c(x, y0) + ψ(y0)) + ε
        if ∃x' ∈ X s.t. σ(x') = y0 then                      ▷ y0 is "stolen" from x'
            S ← S \ {x'}
        S ← S ∪ {x},    σ(x) ← y0
    return σ, ψ
```

Remark 8 (Non-convergence when $\varepsilon = 0$). Consider for instance $X = \{x_1, x_2, x_3\}$, $Y = \{y_1, y_2, y_3\}$, $c(x, y) = \|x - y\|$ and

$$y_1 = (0, 1), \quad y_2 = (0, -1), \quad y_3 = (10, 0),$$
$$x_1 = (-1, 0), \quad x_2 = (-2, 0), \quad x_3 = (-3, 0).$$

The points x_1, x_2 and x_3 are equidistant to y_1 and y_2 and "far" from y_3. Implementing auction's algorithm with $\varepsilon = 0$ then leads to an infinite loop. Indeed, at every steps, the customers x_1, x_2, x_3 pick one of the houses y_1 or y_2, but do not raise the prices, as the second best house is equally interesting. This "bidding war" goes on forever.

Remark 9 (Lower bound on the number of steps). Consider the same setting as before, but with $\varepsilon > 0$. At the beginning of the algorithm, the customers x_1, x_2 and x_3 pick alternatively y_1 or y_2. As long as y_3 has never been selected, the difference of prices between y_1 and y_2 is either 0 or ε, so that the bid is always ε or 2ε. After n iterations, the price of the houses y_1, y_2 is at most equal to $2n\varepsilon$. This means that the third house y_3 will never be chosen until $2n\varepsilon > \min_i \|x_i - y_3\| := C$. As a consequence, the number of iterations is at least $C/(2\varepsilon)$.

The lower bound in the previous remark has a matching upper bound.

Theorem 22. *If one starts the auction algorithm with $\psi = 0$, then*

- *the number of steps in the auction algorithm is at most $N(C/\varepsilon + 1)$, where $C := \max_{X\times Y} c(x, y)$.*
- *the number of operations is at most $N^2(C/\varepsilon + 1)$.*

Moreover, the bijection σ and the prices ψ returned by the algorithm satisfy (CS_ε)*, so that in particular σ is ε-optimal* (3.33)*.*

Remark 10. Note that the computational complexity of this algorithm is very *bad.* Indeed, if $\varepsilon = 10^{-k}$ and if $C = 1$, the number of steps in the worst-case complexity is $10^k N^2$. It would be highly desirable to replace the factor $1/\varepsilon$ by $\log(1/\varepsilon)$. In the next paragraph, we see how this can be achieved using a scaling technique.

The proof of Theorem 22 relies on the following lemma, whose proof is straightforward.

Lemma 23. *Over the course of the auction algorithm,*

(i) the set of selected "houses" $\sigma(S)$ is increasing w.r.t inclusion;
(ii) (ψ, σ) always satisfy the ε-complementary slackness condition ;
(iii) the price increments are by at least ε.

Proof of theorem 22. Suppose that after i steps the algorithm hasn't stopped. Then, there exists a point y_0 in Y that does not belong to $\sigma(S)$, i.e whose price hasn't increased since the beginning of the algorithm, i.e. $\psi(y_0) = 0$.

Suppose now that there exists y_1 whose price has been raised more than $n > C/\epsilon + 1$. Then, by Lemma 23.(iii), one has for every $x \in X$

$$\psi_i(y_0) + c(x, y_0) = c(x, y_0) \leqslant C < n\epsilon - \epsilon \leqslant \psi_i(y_1) - \epsilon \leqslant \psi_i(y_1) + c(x, y_1) - \epsilon$$

This contradicts the fact that y_1 was chosen at a former step. From this, we deduce that there is no point in Y whose price has been raised n times with $n > C/\epsilon + 1$. With at most $C/\varepsilon + 1$ price rise for each of the N objects, and every step costing N (finding the minimum among N) we deduce the desired bound. □

Auction algorithm with ε-scaling. Following [43], Bertsekas and Eckstein [17] modified Algorithm 1 using a scaling technique which improves dramatically both the running time and worst-case complexity of the algorithm. Note that similar scaling techniques have also been applied to improve other algorithms for the assignment problem, see e.g. [54, 47].

The modified algorithm can be described as follows: define $\psi_0 = 0$ and recursively, let ψ_{k+1} be the prices returned by $\mathrm{Auction}(\psi_k, \varepsilon_k)$, where $\varepsilon_k = \frac{C}{2^k}$. One stops when $\varepsilon_k < \varepsilon$, so that the number of runs of the unscaled auction algorithm is bounded by $\log_2(C/\varepsilon)$. Bounding carefully the complexity of each auction run, one gets:

Theorem 24. *The auction algorithm with scaling constructs an η-optimal assignement in time $\mathrm{O}(N^3 \log(C/\eta))$.*

Algorithm 2 Bertsekas-Eckstein auction algorithm with ε-scaling

```
function AuctionScaling(c, η)
    ε ← C, ψ ← 0
    while ε > η do
        σ, ψ ← Auction(c, ε, ψ)
        ε ← ε/2
    return σ, ψ
```

Lemma 25. *Consider a bijection $\sigma_0 : X \to Y$, an injective map $\sigma : S \subseteq X \to Y$, and two price vector $\psi_0, \psi : Y \to \mathbb{R}$. Assume that (σ_0, ψ_0) and (σ, ψ) satisfy respectively the λ- and ε-complementary slackness conditions, with $\varepsilon \leqslant \lambda$. Moreover, suppose that $S \neq X$ and that ψ_0 and ψ agree on the set $Y \setminus \sigma(S)$. Then,*

$$\forall y \in Y,\ \psi(y) \leqslant \psi_0(y) + N(\lambda + \varepsilon)$$

Proof. Consider a point y_0 in Y, and define y_{k+1} as follows: (a) if $y_k \in \sigma(S)$, let $x_k := \sigma^{-1}(y_k)$, and $y_{k+1} = \sigma_0(y_k)$ (b) if $y_k \notin \sigma(S)$, then stop. The ε-complementary slackness for (ψ, σ) at (x_k, y_k) implies

$$c(x_k, y_k)+\psi(y_k) \leqslant \min_{y\in Y} c(x_k, y)+\psi(y)+\varepsilon \leqslant \psi(y_{k+1})+c(x_k, y_{k+1})+\varepsilon. \quad (3.34)$$

Similarly, λ-CS for (ψ_0, σ_0) at (x_k, y_{k+1}) with $y = y_k$ implies

$$\psi_0(y_{k+1}) + c(x_k, y_{k+1}) \leqslant c(x_k, y_k) + \psi_0(y_k) + \lambda. \quad (3.35)$$

Summing the inequalities (3.34) and (3.35) for $k = 0$ to $k = K - 1$ gives

$$\psi_0(y_K) - \psi_0(y_0) + \psi(y_0) - \psi(y_K) \leqslant K \times (\lambda + \varepsilon)$$

By assumption, the point y_K does not belong to $\sigma(S)$ and $\psi(y_K) = \psi_0(y_K)$. This gives us $\psi(y_0) \leqslant \psi_0(y_0)+K(\lambda+\varepsilon)$, and we conclude by remarking that the path $(y_0, x_0, \ldots, y_K)$ is simple, i.e. $K \leqslant N$. □

Proof of Theorem 24. Lemma 25 implies that during the run $k + 1$ of the (unscaled) auction algorithm, the price vector never grows larger than $\psi_0 + (\varepsilon + \lambda)N = \psi_0 + 3\varepsilon N$, with $\lambda := \varepsilon_k$ and $\varepsilon := \frac{1}{2}\lambda$. Since at each step, the price grows by at least ε, there are at most $3N^2$ steps in the run k. Taking into account the cost of finding $\min_{y\in Y} c(x, y) + \psi(y)$ at each step, the computational complexity of each auction run is therefore $O(N^3)$. Since the the number of runs is $O(\log(C/\eta))$, we get the claimed estimate. □

Implementations of auction's algorithm. One of the most expensive phase of Auction's algorithm is the computation of the bid. Computing the bid for a certain customer $x \in X$ requires one to browse through all the houses $y \in Y$ in order to determine the smallest values of $c(x, y) + \psi(y)$, $y \in Y$. The cost of determining the bid accounts for a factor $N = \mathrm{Card}(Y)$ in the computational complexity of auction's algorithm in Theorems 22 and Theorem 24. We mention two possible ways to overcome this difficulty.

Exploiting the geometry of the cost. The first idea is to exploit the geometry of the space in order to reduce the cost of finding the minimum of $c(x,y)+\psi(y)$, $y\in Y$, which accounts for a cost of N in the complexity analysis of Theorem 24. The computation of this minimimum is similar to the *nearest neighbor problem* in computational geometry, and nearest neighbors can sometimes be found in $\log(N)$ time, after some preprocessing. For instance, in the case of $c(x,y)=\|x-y\|^2$ on $\mathbb{R}^d$, and for $\psi\geqslant 0$ one can rewrite

$$c(x,y)+\psi(y)=\|x-y\|^2+(\sqrt{\psi(y)}-0)^2=\left\|(x,0)-(y,\sqrt{\psi(y)})\right\|^2,$$

thus showing that finding the smallest value of $c(x,y)+\psi(y)$ over Y amounts to finding the closest point to $(x,0)$ in the set $\{(y,\sqrt{\psi(y)})\mid y\in Y\}\subseteq\mathbb{R}^{d+1}$. This idea and variants thereof leads to practical and theoretical improvements, both for auction's algorithm and for other algorithms for the assignment problem. We refer to [63, 1] and references therein.

Exploiting the graph structure of solutions. When the cost satisfies the Twist conditition (2.23) on $\mathbb{R}^d$ and the source measure is absolutely continuous, Theorem 12 guarantees that the solution to the Kantorovich's problem is concentrated on a *graph*, i.e. $\dim(\mathrm{spt}(\gamma))=d$ while a priori, the dimension of $\mathrm{spt}(\gamma)\subseteq\mathbb{R}^{2d}$ could be as high as $2d$. It is natural, in view of the stability of the optimal transport plans (Theorem 7), to hope that this feature remains true at the discrete level, meaning that one expects that the support of the discrete solution concentrates on a lower dimensional graph G. One could then try to use this phenomenom to prune the search space, i.e. taking the minimum in $c(x,y)+\psi(y)$ not over the whole space but over points y such that (x,y) lie "close" to G. In practice, G is unknown but can estimated in a coarse-to-fine way. This idea or variants thereof has been used as a heuristic in several works [70, 78, 10], and has been analyzed more precisely by Bernhard Schmitzer [89, 90].

3.3. **Discrete optimal transport via entropic regularization.** We now turn to another method to construct approximate solutions to optimal transport problems between probability measures on two finite sets X and Y. Here, the measures are not supposed uniform any more, and we set

$$\mu=\sum_{x\in X}\mu_x\delta_x\qquad\nu=\sum_{y\in Y}\nu_y\delta_y.$$

For simplicity, we assume throughout that all the points in X and Y carry some mass, that is $\min(\min_{x\in X}\mu_x,\min_{y\in Y}\nu_y)>0$. As before, we conflate a transport plan $\gamma\in\Gamma(\mu,\nu)$ with its density $(\gamma_{x,y})_{(x,y)\in X\times Y}$.

Entropic regularization problem. We start from the primal formulation of the optimal transport problem, but instead of imposing the non-negativity constraints $\gamma_{x,y}\geqslant 0$, we add a term to the transport cost, which penalizes (minus) the entropy of the transport plan and acts as a barrier for the non-negativity constraint:

$$H(\gamma) = \sum_{x\in X, y\in Y} h(\gamma_{x,y}),$$
$$\text{where } h(t) = \begin{cases} t(\log(t)-1) & \text{if } t>0 \\ 0 & \text{if } t=0 \\ +\infty & \text{if } t\leqslant 0 \end{cases} \tag{3.36}$$

The regularized problem is the following minimization problem:

$$(\mathrm{KP}^\eta) := \min_{\gamma\in\overline{\Gamma}(\mu,\nu)} \langle c|\gamma\rangle + \eta H(\gamma),$$
$$\text{where } \overline{\Gamma}(\mu,\nu) = \left\{ \gamma = (\gamma_{x,y}) \mid \sum_{y\in Y} \gamma_{x,y} = \mu_x, \sum_{x\in X} \gamma_{x,y} = \nu_y \right\}. \tag{3.37}$$

Theorem 26. *The problem* (KP^η) *has a unique solution* γ*, which belongs to* $\Gamma(\mu,\nu)$*. Moreover, if* $\min_{x\in X}\mu_x > 0$ *and* $\min_{y\in Y}\mu_y > 0$*, then*

$$\forall (x,y)\in X\times Y,\ \gamma_{x,y} > 0.$$

Lemma 27. $H : \gamma \in (\mathbb{R}_+^*)^{X\times Y} \mapsto \sum_{x,y} h(\gamma_{x,y})$ *is* 1*-strongly convex.*

Proof. From $h''(t) = 1/t$, one sees that the Hessian $\mathrm{D}^2 H(\gamma)$ is diagonal with diagonal coefficients $1/\gamma_{x,y} \geqslant 1$ since $\gamma_{x,y} \in]0,1]$. □

Proof. The regularized problem (KP^η) amounts to minimizing a continuous and coercive function over a closed convex set, thus showing existence. Let us denote by γ^* a solution of (KP^η). Then, γ^* has a finite entropy, so that it satisfies the constraint $\gamma^*_{x,y} \geqslant 0$. This implies that γ^* is a transport map between μ and ν. We now prove by contradiction that the set $Z := \{(x,y) \mid \gamma^*_{x,y} = 0\}$ is empty. For this purpose, we define a new transport map $\gamma^\varepsilon \in \Gamma(\mu,\nu)$ by $\gamma^\varepsilon = (1-\varepsilon)\gamma^* + \varepsilon\mu\otimes\nu$, and we give an upper bound on the energy of γ^ε. We first observe that by convexity of $h : r \mapsto r(\log r - 1)$, one has

$$h(\gamma^\varepsilon_{x,y}) \leqslant (1-\varepsilon)h(\gamma^*_{x,y}) + \varepsilon h(\mu_x\nu_y) \leqslant h(\gamma^*_{x,y}) + O(\varepsilon).$$

We consider some $(x,y)\in Z$. Introducing $C = \min_{x,y}\mu_x\nu_y$, which is strictly positive by assumption, we have

$$\begin{aligned} h(\gamma^\varepsilon_{x,y}) = h(\varepsilon\mu_x\nu_y) &= \mu_x\nu_y\varepsilon(\log\varepsilon + \log(\mu_x\mu_y)) - \mu_x\nu_y\varepsilon \\ &\leqslant C\varepsilon\log\varepsilon + O(\varepsilon), \end{aligned}$$

Summing the two previous estimates over Z and $(X\times Y)\setminus Z$, and setting $n = \mathrm{Card}(Z)$, we get

$$H(\gamma^\varepsilon) \leqslant H(\gamma^*) + Cn\varepsilon\log\varepsilon + O(\varepsilon).$$

Since in addition we have by linearity $\langle c|\gamma^\varepsilon\rangle \leqslant \langle c|\gamma^*\rangle + O(\varepsilon)$, we get

$$\langle c|\gamma^*\rangle + H(\gamma^*) \leqslant \langle c|\gamma^\varepsilon\rangle + H(\gamma^\varepsilon) \leqslant \langle c|\gamma^*\rangle + H(\gamma^*) + Cn\varepsilon\log\varepsilon + O(\varepsilon),$$

where the lower bound comes from the optimality of γ^*. Thus, $Cn\varepsilon\log\varepsilon + O(\varepsilon) \geqslant 0$, which is possible if and only if $n = \mathrm{Card}(Z)$ vanishes, implying the strict positivity of γ^*.

By continuity of the function minimized in (KP$^\eta$), the set of solutions γ^* is closed and therefore included in $[\delta, +\infty)^{X\times Y}$ for some $\delta > 0$. Therefore, by Lemma 27, the regularized problem (KP$^\eta$) amounts to minimizing a coercive and strictly convex function over a closed convex set, thus showing uniqueness of the solution. □

Dual formulation. We start by deriving (formally) the dual problem and first introduce the Lagragian of (KP$^\eta$)

$$L(\gamma, \varphi, \psi) := \sum_{x,y} \gamma_{x,y} c(x, y) + \eta h(\gamma_{x,y}) + \sum_{x\in X} \varphi(x) \left(\mu_x - \sum_{y\in Y} \gamma_{x,y}\right) + \sum_{y\in Y} \psi(y) \left(\sum_{y\in Y} \gamma_{x,y} - \nu_y\right), \tag{3.38}$$

where $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ are the Lagrange multipliers. Then,

$$(\mathrm{KP}^\eta) = \min_\gamma \sup_{\varphi,\psi} L(\gamma, \varphi, \psi).$$

As always, the dual problem is obtained by inverting the infimum and the supremum. We also simplify slightly the expressions:

$$\sup_{\varphi,\psi} \min_\gamma L(\gamma, \varphi, \psi) = \sup_{\varphi,\psi} \min_\gamma \sum_{x,y} \gamma_{x,y}(c(x, y) + \psi(y) - \varphi(x) + \eta(\log(\gamma_{x,y}) - 1)) + \sum_{x\in X} \varphi(x)\mu_x - \sum_{y\in Y} \psi(y)\nu_y. \tag{3.39}$$

Taking the derivative with respect to $\gamma_{x,y}$, we find that for a given φ, ψ, the optimal γ must satisfy:

$$\begin{aligned} &c(x, y) + \psi(y) - \varphi(x) + \eta \log(\gamma_{x,y}) = 0 \\ &\text{i.e. } \gamma_{x,y} = e^{\frac{1}{\eta}(\varphi(x)-\psi(y)-c(x,y))} \end{aligned} \tag{3.40}$$

Putting these values in the Equation (3.39) gives the following definition:

Definition 13 (Dual regularized problem)**.** The dual of the regularized optimal transport problem is defined by

$$(\mathrm{DP}^\eta) = \sup_{\varphi,\psi} \mathcal{K}^\eta(\varphi, \psi) \tag{3.41}$$

where

$$\mathcal{K}^\eta(\varphi, \psi) := - \sum_{(x,y)\in X\times Y} \eta e^{\frac{1}{\eta}(\varphi(x)-\psi(y)-c(x,y))} + \sum_{x\in X} \varphi(x)\mu_x - \sum_{y\in Y} \psi(y)\nu_y. \tag{3.42}$$

We can now state the strong duality result

Theorem 28 (Strong duality)**.** *Strong duality holds and the maximum in the dual problem is reached, i.e. there exist $\varphi \in \mathbb{R}^X$ an $\psi \in \mathbb{R}^Y$ such that*

$$(\mathrm{KP}^\eta) = (\mathrm{DP}^\eta) = \mathcal{K}^\eta(\varphi, \psi).$$

Corollary 29. *If φ, ψ is the solution to the dual problem* (DP$^\eta$)*, then the solution γ of* (KP$^\eta$) *is given by*

$$\gamma_{x,y} = e^{\frac{\varphi(x)-\psi(y)-c(x,y)}{\eta}}.$$

Corollary 29 is a direct consequence of the relation (3.40). This holds because, unlike the original linear programming formulation of optimal transport, the regularized problem (KP$^\eta$) is smooth and strictly convex.

Proof of Theorem 28. Weak duality (KP$^\eta$) $\geqslant$ (DP$^\eta$) always hold. To prove the strong duality, we denote by γ^* the solution to (KP$^\eta$), and we note that by Theorem 26, $\gamma^*_{xy} > 0$ for all $(x, y) \in X \times Y$. This implies that the optimized functional $\gamma \mapsto \langle c|\gamma\rangle + \eta H(\gamma)$ is $\mathcal{C}^1$ in a neighborhood of γ^*. Thus, there exists Lagrange multipliers for the equality constrained problem, i.e. $\tilde{\varphi} \in \mathbb{R}^X$ and $\tilde{\psi} \in \mathbb{R}^Y$ such that

$$\nabla_\gamma L(\gamma^*, \tilde{\varphi}, \tilde{\psi}) = 0.$$

Since the function $L(\cdot, \tilde{\varphi}, \tilde{\psi})$ is convex, this implies that $\gamma^* = \mathrm{argmin}_\gamma L(\gamma, \tilde{\varphi}, \tilde{\psi})$. Hence

$$(\mathrm{DP}^\eta) = \sup_{\varphi,\psi} \min_\gamma L(\gamma, \varphi, \psi) \geqslant \min_\gamma L(\gamma, \tilde{\varphi}, \tilde{\psi}) = L(\gamma^*, \tilde{\varphi}, \tilde{\psi}) = (\mathrm{KP}^\eta).$$

The last equality follows from the fact that γ^* satisfies the constraints and is a solution to (KP$^\eta$). Thus (DP$^\eta$) = (KP$^\eta$). □

Regularized c-transform. A natural way to maximize $\mathcal{K}^\eta(\varphi, \psi)$ is to maximize alternatively in φ and ψ. In the case of entropy-regularized optimal transport, each of the partial maximization problems ($\max_\varphi \mathcal{K}^\eta(\varphi, \psi)$ and $\max_\psi \mathcal{K}^\eta(\varphi, \psi)$) have explicit solutions, which are connected to the notion of c-transform in (non-regularized) optimal transport:

Proposition 30. *The following holds*

(i) *Given $\psi \in \mathbb{R}^Y$, the maximizer of $\mathcal{K}^\eta(\cdot, \psi)$ is attained at a unique point in $\mathbb{R}^X$, denoted $\psi^{c,\eta}$, and defined by*

$$\psi^{c,\eta}(x) = \eta \log(\mu_x) - \eta \log\left(\sum_{y \in Y} e^{\frac{1}{\eta}(-c(x,y)-\psi(y))}\right). \tag{3.43}$$

(ii) *Given, $\varphi \in \mathbb{R}^X$, the maximizer of $\mathcal{K}^\eta(\varphi, \cdot)$ is attained at a unique point in $\mathbb{R}^Y$, denoted $\varphi^{\bar{c},\eta}$, and defined by*

$$\varphi^{\bar{c},\eta}(y) = -\eta \log(\mu_y) + \eta \log\left(\sum_{x \in X} e^{\frac{1}{\eta}(-c(x,y)+\varphi(x))}\right). \tag{3.44}$$

Proof. To prove (i), consider $\varphi \in \mathbb{R}^X$ the maximizer of $\mathcal{K}^\eta(\cdot, \psi)$. Taking the derivative of $\mathcal{K}^\eta$ with respect to the variable $\varphi(x)$ gives us

$$\mu_x = e^{\frac{\varphi(x)}{\eta}} \sum_{y \in Y} e^{-\frac{1}{\eta}(\psi(y)+c(x,y))},$$

implying the desired formula. The second formula is proven similarly. □

Definition 14 (Regularized c-transform)**.** Given $\psi \in \mathbb{R}^Y$, we will call the function $\psi^{c,\eta}$ defined by (3.43) its *regularized c-transform.* Similarly, given $\varphi \in \mathbb{R}^X$, we call the function $\varphi^{\overline{c},\eta}$ defined by (3.44) its *regularized $\overline{c}$-transform*

Remark 11 (Relation to the c-transform). As the notation indicates, $\psi^{c,\eta}$ is related to the c-transform used in optimal transport (Def. 7). Indeed, when η tends to zero, one has

$$\begin{aligned}\lim_{\eta\to 0} \psi^{c,\eta}(x) &= \lim_{\eta\to 0} \eta\left(\log(\mu_x) - \log\left(\sum_{y\in Y} e^{\frac{1}{\eta}(-c(x,y)-\psi(y))}\right)\right)\\ &= \min_{y\in Y} c(x,y) + \psi(y) = \psi^c(x).\end{aligned}$$

This explains the choice of notation: $\psi^{c,\eta}$ is a smoothed version of the c-transform introduced in Definition 7.

The following two properties are very similar to some properties holding for the standard c-transform. In the following, we denote $\|\cdot\|_{o,\infty}$ the pseudo-norm of uniform convergence up to addition of a constant:

$$\|f\|_{o,\infty} = \inf_{a\in\mathbb{R}} \|f+a\|_\infty = \frac{1}{2}(\sup f - \inf f).$$

This pseudo-norm will be very useful to state convergence results for Sinkhorn-Knopp's algorithm for solving the regularized optimal transport problem.

Proposition 31. *Let $\psi, \overline{\psi} \in \mathbb{R}^Y$. Then,*

(i) for $a \in \mathbb{R}$, $(\psi + a)^{c,\eta} = \psi^{c,\eta} + a$.
(ii) $\|\psi^{c,\eta}\|_{o,\infty} \leqslant \eta\,\|\log(\nu)\|_{o,\infty} + \|c\|_{o,\infty}$,
(iii) $\left\|\psi^{c,\eta} - \overline{\psi}^{c,\eta}\right\|_{o,\infty} \leqslant \|\psi - \overline{\psi}\|_{o,\infty}$.

Similar properties hold for the map $\varphi \in \mathbb{R}^X \mapsto \varphi^{\overline{c},\eta}$.

Proof. (ii) Using the formula (3.43), and $c(x,y) - c(x',y) \leqslant \sup c - \inf c$,

$$\begin{aligned}&\psi^{c,\eta}(x) - \psi^{c,\eta}(x')\\ &\quad= \eta(\log(\mu_x) - \log(\mu_{x'}))\\ &\qquad+ \eta\left(\log\left(\sum_{y\in Y} e^{\frac{1}{\eta}(-c(x',y)-\psi(y))}\right) - \log\left(\sum_{y\in Y} e^{\frac{1}{\eta}(-c(x,y)-\psi(y))}\right)\right).\\ &\quad\leqslant \eta(\sup\log(\mu) - \inf\log(\mu)) + \sup c - \inf c,\end{aligned}$$

implying the first inequality.

(iii) If we show that $\left\|\psi^{c,\eta} - \overline{\psi}^{c,\eta}\right\|_\infty \leqslant \|\psi - \overline{\psi}\|_\infty$, the same inequality with $\|\cdot\|_{o,\infty}$ will follow easily using (i). Using $\psi(y) \leqslant \overline{\psi}(y) + \|\psi - \overline{\psi}\|_\infty$, we have

$$\begin{aligned}&\psi^{c,\eta}(x) - \overline{\psi}^{c,\eta}(x)\\ &\quad= -\eta\log\left(\sum_{y\in Y} e^{\frac{1}{\eta}(-c(x,y)-\psi(y))}\right) + \eta\log\left(\sum_{y\in Y} e^{\frac{1}{\eta}(-c(x,y)-\overline{\psi}(y))}\right)\\ &\quad\leqslant \|\psi - \overline{\psi}\|_\infty\end{aligned}$$

□

Regularized Kantorovitch functional. As in standard optimal transport (see §2.3) and following Cuturi and Peyré [36], we can express the regularized dual maximization problem (3.41) using only the variable $\psi \in \mathbb{R}^Y$.

Definition 15 (Regularized Kantorovitch functional)**.** The regularized Kantorovitch functional $\mathcal{K}^\eta : \mathbb{R}^Y \to \mathbb{R}$ is defined by

$$\mathcal{K}^\eta(\psi) = \max_{\varphi \in \mathbb{R}^X} \mathcal{K}^\eta(\varphi, \psi) = \langle \psi^{c,\eta} | \mu \rangle - \langle \psi | \nu \rangle \tag{3.45}$$

Since $\psi^{c,\eta}$ has a closed-form expression, the functional $\mathcal{K}^\eta$ can be computed explicitely. This explicit expression is a special feature of the choice of the entropy as the regularization. In the next formula, $H(\mu) = \sum_{x \in X} \mu_x \log(\mu_x)$:

$$\mathcal{K}^\eta(\psi) = -\eta \sum_{x \in X} \mu_x \left(\log \sum_{y \in Y} e^{\frac{-c(x,y) - \psi(y)}{\eta}} \right) + \eta H(\mu) - \sum_{y \in Y} \psi(y) \nu_y,$$

Remark 12. Note the similarity between the formula for Kantorovich functional derived from regularized transport (3.45) and the formula for the Kantorovich functional without regularization (2.24). Note also that $\mathcal{K}^\eta$ is also invariant by addition of a constant, namely $\mathcal{K}^\eta(\psi + \lambda \mathbf{1}_Y) = \mathcal{K}^\eta(\psi)$ for any $\lambda \in \mathbb{R}$ and $\mathbf{1}_Y = \sum_{y \in Y} \mathbf{1}_y$ the constant function equal to one.

In order to express the gradient and the Hessian of $\mathcal{K}^\eta$, we introduce the notion of *smoothed laguerre cells.*

Definition 16 (Smoothed Laguerre cells)**.** Given $\psi \in \mathbb{R}^Y$, we define

$$\mathrm{RLag}_y^\eta(\psi) = \frac{e^{-\frac{c(\cdot,y)+\psi(y)}{\eta}}}{\sum_{z \in Y} e^{-\frac{c(\cdot,z)+\psi(z)}{\eta}}}. \tag{3.46}$$

Unlike the standard Laguerre cell $\mathrm{Lag}_y(\psi)$ defined in (2.27), which is a set, $\mathrm{RLag}_y^\eta(\psi)$ is a *function.* The family $(\mathrm{RLag}_y^\eta(\psi))_{y \in Y}$ is a partition of unity, meaning that the sum over y of $\mathrm{RLag}_y^\eta(\psi)$ equals one. One can loosely think of the regularized Laguerre cells as smoothed indicator functions of the (standard) Laguerre cells. In particular,

$$\lim_{\eta \to 0} \mathrm{RLag}_y^\eta(\psi)(x) = \begin{cases} 0 & \text{if } x \notin \mathrm{Lag}_y(\psi) \\ 1 & \text{if } x \in \mathrm{SLag}_y(\psi), \end{cases}$$

where $\mathrm{SLag}_y(\psi)$ is the strict Laguerre cell introduced in (2.28). We also introduce the two quantities

$$G_y^\eta(\psi) = \langle \mathrm{RLag}_y^\eta(\psi) | \mu \rangle$$

$$G_{yz}^\eta(\psi) = \begin{cases} \frac{1}{\eta} \langle \mathrm{RLag}_y^\eta(\psi) \mathrm{RLag}_z^\eta(\psi) | \mu \rangle & \text{if } z \neq y \\ -\sum_{z \neq y} G_{yz}^\eta(\psi) & \text{if } z = y. \end{cases}$$

Informally, $G_y^\eta(\psi)$ measures the quantity of mass of μ within the regularized Laguerre cell $\mathrm{RLag}_y^\eta(\psi)$.

Theorem 32.
• *The regularized Kantorovitch functional $\mathcal{K}^\eta$ is $\mathcal{C}^\infty$, concave, with first and second-order partial derivatives given by*

$$\forall y \in Y,\ \frac{\partial \mathcal{K}^\eta}{\partial \mathbf{1}_y}(\psi) = G^\eta_y(\psi) - \nu_y,$$

$$\forall y \neq z \in Y,\ \frac{\partial^2 \mathcal{K}^\eta}{\partial \mathbf{1}_z \partial \mathbf{1}_y}(\psi) = G^\eta_{yz}(\psi).$$

• *The function $\mathcal{K}^\eta$ is strictly concave on the orthogonal of the set of constant functions. More precisely, for every $\psi \in \mathbb{R}^Y$ one has*

$$\forall v \in \mathbb{R}^Y \text{ s.t.} \sum_{y\in Y} v(y) = 0,\ \mathrm{D}^2\mathcal{K}^\eta(\psi)(v,v) < 0.$$

• *If ψ is a maximizer in* (DP$^\eta$), *then the solution to* (KP$^\eta$) *is given by*

$$\gamma = \sum_{x,y} \gamma_{x,y}\delta_{(x,y)},\ \text{with } \gamma_{x,y} = \mathrm{RLag}^\eta_y(\psi)(x)\mu_x.$$

Proof. For every $y \in Y$, the derivative is given by

$$\frac{\partial \mathcal{K}^\eta}{\partial \mathbf{1}_y}(\psi) = \sum_{x\in X} \mu_x \frac{e^{\frac{-\psi(y)-c(x,y)}{\eta}}}{\sum_{z\in Y} e^{\frac{-c(x,z)-\psi(z)}{\eta}}} - \nu_y = G^\eta_y(\psi) - \nu_y.$$

The second order derivative is given for $z \neq y$ by

$$\frac{\partial^2 \mathcal{K}^\eta}{\partial \mathbf{1}_z \partial \mathbf{1}_y}(\psi) = \sum_{x\in X} \mu_x e^{\frac{-\psi(y)-c(x,y)}{\eta}} \frac{\frac{1}{\eta} e^{\frac{-\psi(z)-c(x,z)}{\eta}}}{\left(\sum_{z\in Y} e^{\frac{-c(x,z)-\psi(z)}{\eta}}\right)^2} = G^\eta_{yz}(\psi).$$

The relation

$$\sum_{y\in Y} \frac{\partial \mathcal{K}^\eta}{\partial \mathbf{1}_y}(\psi) = 1$$

gives the desired formula for the second order derivatives when $z = y$. The hessian of $\mathcal{K}^n$ is therefore symmetric with dominant diagonal, with negative diagonal coefficients. This implies that the Hessian is negative, hence that $\mathcal{K}^\eta$ is concave. Let us now show that $\ker H = \mathbb{R}\mathbf{1}_Y$, where $H = \mathrm{D}^2\mathcal{K}^\eta(\psi)$. Consider $v \in \ker H$ and let $y_0 \in Y$ be the point where v attains its maximum. Then using $Hv = 0$, and in particular $(Hv)(y_0) = 0$, one has

$$\begin{aligned} 0 &= \left(\sum_{y\neq y_0} H_{y,y_0} v(y)\right) + H_{y_0,y_0} v(y_0) \\ &= \sum_{y\neq y_0} H_{y,y_0}(v(y) - v(y_0)). \end{aligned}$$

This follows from $H_{y_0,y_0} = -\sum_{y\neq y_0} H_{y,y_0}$. Since for every $y \neq y_0$, one has $H_{y,y_0} > 0$ and $v(y_0) - v(y) \geqslant 0$, this implies that $v(y) = v(y_0)$. Therefore $\ker H \subseteq \mathbb{R}\mathbf{1}_Y$. The reverse inclusion is obvious and therefore $\mathcal{K}^\eta$ is strictly concave on the orthogonal of the set of constant functions.

To prove the last claim we note that if ψ maximizes $\mathcal{K}^\eta(\cdot)$, then $(\psi^{c,\eta},\psi)$ maximizes $\mathcal{K}^\eta(\cdot,\cdot)$. By Corollary 29, the optimal transport map γ is

$$\gamma_{x,y} = e^{\frac{\varphi(x)-\psi(y)-c(x,y)}{\eta}} = \frac{e^{-\frac{\psi(y)+c(x,y)}{\eta}}\mu_x}{\sum_{z\in Y} e^{-\frac{\psi(z)+c(x,z)}{\eta}}} = \mathrm{RLag}_y^\eta(\psi)(x)\mu_x. \qquad \square$$

Sinkhorn-Knopp as block coordinate ascent. We present here the Sinkhorn-Knopp algorithm that consists in computing a maximizer to the dual problem (DP$^\eta$) by optimizing the functional $\mathcal{K}^\eta$ alternatively in φ and ψ. The iterations are defined by

$$\begin{cases} \varphi^{(k+1)} = (\psi^{(k)})^{c,\eta} \\ \psi^{(k+1)} = (\varphi^{(k+1)})^{\bar{c},\eta}, \end{cases} \tag{3.47}$$

or equivalently $\psi^{(k+1)} = S(\psi^{(k)})$ where

$$S(\psi) = (\psi^{c,\eta})^{\bar{c},\eta}. \tag{3.48}$$

Remark 13 (Relation to matrix factorization). This algorithm is in fact a reformulation, using a logarithmic change of variable, of Sinkhorn-Knopp's algorithm [92] for finding a factorization of non-negative matrices introduced by Sinkhorn [91]. We therefore refer to the iterations (3.47)–(3.48) as Sinkhorn-Knopp's algorithm.

Correctness. We first show the correctness of Sinkhorn–Knopp's algorithm, using a simple expression for $S(\psi)$ which can be found in an article of Robert Berman [14].

Proposition 33 (Correctness of Sinkhorn-Knopp)**.** *Let $\psi\in\mathbb{R}^Y$ be a potential. The following assertions are equivalent:*

(i) ψ is a fixed point of S;
(ii) for every $y\in Y$ $\langle\mu|\mathrm{RLag}_y^\eta(\psi)\rangle = \nu_y$;
(iii) ψ is a maximizer of the regularized Kantorovich function $\mathcal{K}^\eta$

This proposition follows at once from the next lemma, and from the computation of $\nabla\mathcal{K}^\eta$ in Theorem 32.

Lemma 34. $\frac{S(\psi)(y)-\psi(y)}{\eta} = -\log(\nu_y) + \log\langle\mu|\mathrm{RLag}_y^\eta(\psi)\rangle$.

Proof. A calculation shows that

$$\begin{aligned} S(\psi)(y) &= -\eta\left(\log(\nu_y) - \log\sum_{x\in X} e^{\frac{-c(x,y)+\eta\left(\log(\mu_x)-\log\sum_{z\in Z} e^{\frac{-c(z,y)-\psi(z)}{\eta}}\right)}{\eta}}\right) \\ &= -\eta\left(\log(\nu_y) - \log\sum_{x\in X}\mu_x\frac{e^{\frac{-c(x,y)}{\eta}}}{\sum_{z\in Y} e^{\frac{-c(z,y)-\psi(z)}{\eta}}}\right) \\ &= -\eta\left(\log(\nu_y) - \log e^{\frac{\psi(y)}{\eta}}\sum_{x\in X}\mu_x\mathrm{RLag}_y^\eta(\psi)(x)\right) \\ &= -\eta\left(\log(\nu_y) - \log e^{\frac{\psi(y)}{\eta}}\langle\mu|\mathrm{RLag}_y^\eta(\psi)\rangle\right), \end{aligned}$$

which implies the equation. □

Convergence. In order to prove convergence, we need to strengthen the 1-Lipschitz estimation from Proposition 31. This allows to apply Picard's fixed point theorem to get the contraction of the Sinkhorn-Knopp iteration (3.48). The proof we present in this chapter has been first introduced in course notes of Vialard [96].

Theorem 35 (Convergence of Sinkhorn, [96]). *The map S is a contraction for $\|\cdot\|_{o,\infty}$. More precisely,*

$$\left\|S(\psi^0) - S(\psi^1)\right\|_{o,\infty} \leqslant \left(1 - e^{-2\frac{\|c\|_{o,\infty}}{\eta}}\right) \left\|\psi^0 - \psi^1\right\|_{o,\infty}.$$

In particular, the iterates $(\varphi^{(k)}, \psi^{(k)})$ of Sinkhorn-Knopp's algorithm (3.47) *converge with linear rate to the unique (up to constant) maximizer the regularized dual problem* (3.41)

Remark 14 (Other convergence proofs). The convergence of Sinkhorn-Knopp's algorithm is usually proven (e.g. in [92]) using a theorem of Birkhoff [18]. We refer to the recent book by Peyré and Cuturi [83] for this point of view. Other convergence proofs exist, see for instance Berman [14] (in the continuous case), and Altschuler, Weed and Rigolet [3].

Remark 15 (Convergence speed). This theorem shows that the Sinkhorn-Knopp algorithm converges with linear speed, but the contraction constant has a bad dependency in η. Denoting $C = \|c\|_{o,\infty}$, to get an error of ε one needs

$$(1 - e^{-2C/\eta})^k \leqslant \varepsilon$$
$$\text{i.e. } k \gtrsim e^{2C/\eta} \log(1/\varepsilon),$$

where the second inequality holds for small values of η. This bad dependency in η seems to be a practical obstacle to choosing a very small smoothing parameter. This calls for scaling techniques, as for the auction's algorithm, and was considered by Schmitzer [89, 90].

Remark 16 (Implementation). The numerical implementation of Sinkhorn-Knopp's algorithm is more complicated than it seems:

- In a naive implementation, the computation of the smoothed c-transforms (3.43)–(3.44) has a cost proportional to $\mathrm{Card}(X)\mathrm{Card}(Y)$. This can be alleviated for instance when $X = Y$ are grids and when the cost is a $\|\cdot\|_p$ norm, using fast convolution techniques (see e.g. [93] or [83, Remark 4.17]), or when the cost is the squared geodesic distance on a Riemannian manifold [33, 93].
- The convergence speed can be slow when the supports of the data X, Y are "far" from each other, and when η is small. This difficulty is cirvumvented using the η-scaling techniques mentioned above, often combined with multi-scale (coarse-to-fine) strategies, studied in this context by Benamou, Carlier and Nenna [11] and Schmitzer [89].
- Finally, some numerical difficulties (divisions by zero) can occur when η is small and the potential ψ is far from the solution.

The book of Cuturi and Peyré present these difficulties in more details and explain how to circumvent them [83]. In addition to the works already cited, we refer to the PhD work of Feydy [29, 45], and especially to the implementation of regularized optimal transport in the library GeomLoss[2].

In order to prove this theorem, we will make use of the following elementary lemma, giving an upper bound on the L^1 distance between two Gibbs kernels e^{u_i}/Z_i for $i \in \{0,1\}$ as a function of $\|u_1 - u_0\|_{o,\infty}$.

Lemma 36. *Let u_0, u_1 be two functions on Y and denote $g_i = e^{u_i}/Z_i$ where $Z_i = \sum_{y\in Y} e^{u_i(y)}$. Then,*

$$\sum_{y\in Y} |g_1(y) - g_0(y)| \leqslant 2(1 - e^{-2\|u_0-u_1\|_{o,\infty}}).$$

Proof. Note that by definition the Gibbs kernel g_i does not change if a constant is added to u_i, so that we can assume that

$$\varepsilon := \|u_0 - u_1\|_{o,\infty} = \|u_0 - u_1\|_\infty .$$

Using the inequality $u_0 - \varepsilon \leqslant u_1 \leqslant u_0 + \varepsilon$, one easily shows that

$$e^{-2\varepsilon}\frac{e^{u_0}}{Z_0} \leqslant \frac{u_1}{Z_1} \leqslant e^{2\varepsilon}\frac{e^{u_0}}{Z_0},$$

thus implying $e^{-2\varepsilon} g_0 \leqslant g_1 \leqslant e^{2\varepsilon} g_0$. This gives

$$\begin{cases}(e^{-2\varepsilon} - 1)g_0 \leqslant g_1 - g_0\\ (e^{-2\varepsilon} - 1)g_1 \leqslant g_0 - g_1,\end{cases}$$

thus implying

$$|g_1 - g_0| \leqslant (1 - e^{-2\varepsilon}) \max(g_0, g_1) \leqslant (1 - e^{-2\varepsilon})(g_0 + g_1).$$

Summing this inequality over Y and using $\sum_Y g_i = 1$, we obtain the desired inequality. □

Proof of Theorem 35. Consider $\psi_0, \psi_1 \in \mathbb{R}^Y$ and $\psi_t = \psi_0 + tv$ with $v = \psi_1 - \psi_0$. Without loss of generality, we assume that the functions ψ_0, ψ_1 are translated by a constant so that $\|\psi_0 - \psi_1\|_\infty = \|\psi_0 - \psi_1\|_{o,\infty}$. We will first give an upper bound on $\|\psi_1^{c,\eta} - \psi_0^{c,\eta}\|_{o,\infty}$, and to do that we will give an upper bound on

$$A(x, x') = (\psi_1^{c,\eta}(x) - \psi_0^{c,\eta}(x)) - (\psi_1^{c,\eta}(x') - \psi_0^{c,\eta}(x'))$$

which is independent of $x, x' \in X$. For this purpose, we introduce

$$B(t, x, x') = -\eta \log\left(\sum_{y\in Y} e^{\frac{1}{\eta}(-c(x,y)-\psi_t(y))}\right) + \eta \log\left(\sum_{y\in Y} e^{\frac{1}{\eta}(-c(x',y)-\psi_t(y))}\right),$$

and

$$g_{x,t}(y) = \frac{e^{\frac{1}{\eta}(-c(x,y)-\psi_t(y))}}{\sum_{z\in Y} e^{\frac{1}{\eta}(-c(x,z)-\psi_t(z))}}.$$

[2] https://www.kernel-operations.io/geomloss/

Then, recalling the definition of $\psi_t^{c,\eta}$ in Eq. (3.44),

$$\begin{aligned}A(x,x') &= B(1,x,x') - B(0,x,x')\\ &= \int_0^1 \partial_t B(t,x,x')\mathrm{d}t = \int_0^1 \langle v | g_{x,t} - g_{x',t}\rangle_{\mathbb{R}^Y}\mathrm{d}t,\\ &\leqslant \|v\|_\infty \int_0^1 \sum_{y\in Y} \left|g_{x,t}(y) - g_{x',t}(y)\right| \mathrm{d}t\end{aligned}$$

Then, by the previous lemma (Lemma 36) and setting $u_{x,t}(y) = -\frac{1}{\eta}(c(x,y)+\psi_t(y))$, so that $g_{x,t} = e^{u_{x,t}}/Z_{x,t}$ with $Z_{x,t} = \sum_Y g_{x,t}$, we obtain

$$\begin{aligned}A(x,x') &\leqslant 2\|v\|_\infty \int_0^1 1 - e^{-2\left\|u_{x,t}-u_{x',t}\right\|_{o,\infty}}\mathrm{d}t\\ &\leqslant 2\|\psi_1 - \psi_0\|_{o,\infty}(1 - e^{-2\left\|u_{x,t}-u_{x',t}\right\|_{o,\infty}})\end{aligned}$$

In addition,

$$\left\|u_{x,t} - u_{x',t}\right\|_{o,\infty} \leqslant \frac{\|c\|_{o,\infty}}{\eta}.$$

We therefore obtain

$$\begin{aligned}\|\psi_1^{c,\eta} - \psi_0^{c,\eta}\|_{o,\infty} &\leqslant \frac{1}{2}\sup_{x,x'\in X} A(x,x')\\ &\leqslant \|\psi_1 - \psi_0\|_{o,\infty}\left(1 - e^{-2\frac{\|c\|_{o,\infty}}{\eta}}\right).\end{aligned}$$

We conclude the proof of the contraction inequality by remarking that the map $\varphi \mapsto \varphi^{\bar{c},\eta}$ is 1-Lipschitz, thanks to Proposition 31.(iii). □

4. Semi-discrete optimal transport

In this part, we consider the semi-discrete optimal transport problem, where the source measure is a probability density and the target is a finitely supported measure. We start by introducing in Section 4.1 the framework of semi-discrete optimal transport, showing its connection with the notion of Laguerre tessellation in discrete geometry. We study in detail the regularity of Kantorovitch functional $\mathcal{K}$ in this setting, in connection with algorithms for solving the semi-discrete optimal transport problem:

- In Section 4.2, we show convergence of the *coordinate-wise increment* algorithm introduced by Oliker and Prüssner using the Lipschiz-continuity of the gradient $\nabla\mathcal{K}$.
- In Section 4.3 a damped Newton method and prove its convergence from a C^1-regularity and monotonicity property of the gradient $\nabla\mathcal{K}$.
- Finally, we consider in Section 4.4 the entropic regularization of the semi-discrete optimal transport problem and its relation to unregularized semi-discrete optimal transport.

4.1. Formulation of semi-discrete optimal transport.

Formulation of semi-discrete optimal transport. Our working assumptions for this section are the following:

- Ω_X, Ω_Y are two open subsets of $\mathbb{R}^d$. The cost function $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$ satisfies the twist condition introduced in Definition 8.

- the source measure ρ is absolutely continuous with respect to the Lebesgue measure on Ω_X and its support is contained in a compact subset X of Ω_X. **When writing $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$ we always mean that ρ belongs to $\mathcal{P}^{\mathrm{ac}}(\Omega_X)$ with $\mathrm{spt}(\rho) \subseteq X$.**
- the target space Y is finite so that $\nu \in \mathcal{P}(Y)$ can be written under the form $\nu = \sum_{y \in Y} \nu_y \delta_y$. For simplicity, we assume that $\min_y \nu_y > 0$.

Note that by an abuse of notation, we will often conflate ρ with its density with respect to the Lebesgue measure.

Laguerre tessellation. In the semi-discrete setting, the dual of Kantorovich's relaxation can be conveniently phrased using the notion of Laguerre tessellation, a variant of the Voronoi tesselation. This connection was already known and used in the 1980s and 1990s, see for instance Cullen–Purser [34], Aurenhammer–Hoffman–Aronov [6] or Gangbo-McCann [51], Caffarelli–Kochengin–Oliker [26]. Large-scale numerical implementations are more recent, starting in the 2010s, see e.g. [70, 38, 55, 64, 68, 65, 42, 58, 40, 41]. To explain the connection, we start with an economic metaphor. Assume that the probability density ρ describes the population distribution over a large city Ω_X, and that the finite set Y describes the location of bakeries in the city. Customers living at a location x in Ω_X try to minimize the walking cost $c(x, y)$, resulting in a decomposition of the space called a Voronoi tessellation. The number of customers received by a bakery $y \in Y$ is equal to the integral of ρ over its Voronoi cell,

$$\mathrm{Vor}_y := \{x \in \Omega_X \mid \forall z \in Y, c(x, y) \leqslant c(x, z)\}.$$

If the price of bread is given by a function $\psi : Y \to \mathbb{R}$, customers living at location x in X make a compromise between walking cost and price by minimizing the sum $c(x, y) + \psi(y)$. This leads to the notion of Laguerre tessellation.

Definition 17 (Laguerre tessellation)**.** The Laguerre tessellation associated to a set of prices $\psi : Y \to \mathbb{R}$ is a decomposition of the space into *Laguerre cells* defined by

$$\mathrm{Lag}_y(\psi) := \{x \in \Omega_X \mid \forall z \in Y, c(x, y) + \psi(y) \leqslant c(x, z) + \psi(z)\}. \tag{4.49}$$

More generally, for any distinct $y_1, \dots, y_\ell \in Y$, we denote the common facet between the Laguerre cells $\mathrm{Lag}_{y_i}(\psi)$ by

$$\mathrm{Lag}_{y_1 \dots y_\ell}(\psi) = \bigcap_{1 \leqslant i \leqslant \ell} \mathrm{Lag}_{y_i}(\psi) \tag{4.50}$$

We will also frequently consider the following hypersurfaces/halfspaces:

$$\begin{aligned} H_{yz}(\psi) &= \{x \in \Omega_X \mid c(x, y) + \psi(y) = c(x, z) + \psi(z)\}, \\ H_{yz}^{\leqslant}(\psi) &= \{x \in \Omega_X \mid c(x, y) + \psi(y) \leqslant c(x, z) + \psi(z)\}, \end{aligned} \tag{4.51}$$

which are defined so that

$$\begin{aligned} &\forall y \in Y, \mathrm{Lag}_y(\psi) = \cap_{z \in Y \setminus \{y\}} H_{yz}^{\leqslant}(\psi) \\ &\forall y, z \in Y, \mathrm{Lag}_{yz}(\psi) \subseteq H_{yz}(\psi) \end{aligned} \tag{4.52}$$

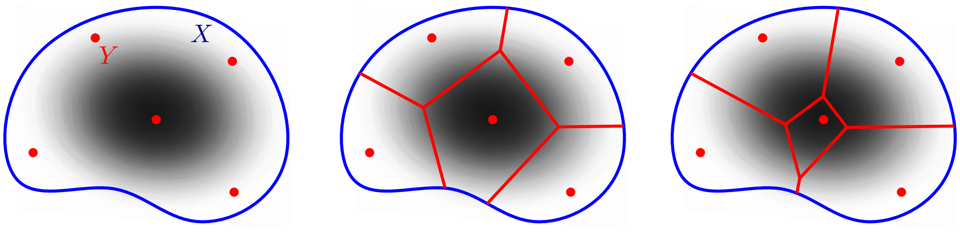

Figure 1. (Left) The domain X (with boundary in blue) is endowed with a probability density pictured in grayscale representing the density of population in a city. The set Y (in red) represents the location of bakeries. Here, $X, Y \subseteq \mathbb{R}^2$ and $c(x, y) = |x - y|^2$ (Middle) The Voronoi tessellation induced by the bakeries (Right) The Laguerre tessellation: the price of bread the bakery near the center of X is higher than at the other bakeries, effectively shrinking its Laguerre cell.

Remark 17. For the quadratic cost $c(x, y) = \|x - y\|^2$, one has

$$\begin{aligned} & c(x, y) + \psi(y) \leqslant c(x, z) + \psi(z) \\ \iff & \langle x|z - y\rangle \leqslant \frac{1}{2}(\psi(z) + \|z\|^2 - (\psi(y) - \|y\|^2)), \end{aligned}$$

which easily implies that the Laguerre cells are convex polyhedra intersected with the domain Ω_X. Introducing $\tilde{\psi}(z) = \frac{1}{2}(\psi(z) + \|z\|^2)$, one has

$$\mathrm{Lag}_y(\psi) = \{x \in \Omega_X \mid \forall z \in Y,\ \langle x|z - y\rangle \leqslant \tilde{\psi}(z) - \tilde{\psi}(y)\}.$$

As a direct consequence, the intersection of two distinct Laguerre cells is contained in an hyperplane and is therefore Lebesgue negligible. If in addition $\psi \equiv 0$, then the Laguerre tessellation coincides with the Voronoi tessellation. The shape of the Voronoi and Laguerre tessellations is depicted in Figure 1.

The following proposition shows that Laguerre tessellations can be used to build optimal transport maps.

Proposition 37. *Under the twist condition (Def. 8), the intersection of two distinct Laguerre cells* $\mathrm{Lag}_y(\psi) \cap \mathrm{Lag}_z(\psi)$ *($y \neq z$) is Lebesgue-negligible, and the map*

$$T_\psi : x \in \Omega_X \mapsto \arg\min_{y \in Y} c(x, y) + \psi(y)$$

is well-defined Lebesgue almost-everywhere. In addition for any $\psi \in \mathbb{R}^Y$ *and any* $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$*,* T_ψ *is an optimal transport map for the cost* c *between* ρ *and the measure*

$$\nu_\psi := T_{\psi\#}\rho = \sum_{y \in Y} \rho(\mathrm{Lag}_y(\psi))\delta_y. \tag{4.53}$$

Proof. By Equation (4.52), one has

$$\mathrm{Lag}_y(\psi) \cap \mathrm{Lag}_z(\psi) = \mathrm{Lag}_{yz}(\psi) \subseteq f^{-1}(\{0\}),$$

where we have set $f(x) = c(x, y) - c(x, z) + \psi(y) - \psi(z)$. By the twist condition, $\nabla f(x) \neq 0$ for all $x \in \Omega_X$, implying that the set $f^{-1}(\{0\})$ is a $(d-1)$-submanifold and is in particular Lebesgue-negligible. This easily implies that T_ψ is well-defined.

Let us now prove optimality of T_ψ in the optimal transport problem between ρ and $T_{\psi\#}\rho$. By definition of T_ψ, one has

$$\forall (x,y) \in X,\ c(x, T_\psi(x)) + \psi(T_\psi(x)) \leqslant c(x,y) + \psi(y).$$

Let γ be a transport plan between ρ and ν_ψ. Integrating the above inequality with respect to γ gives

$$\int_X (c(x, T_\psi(x)) + \psi(T_\psi(x)))\rho(x)\mathrm{d}x \leqslant \int_{X\times Y} (c(x,y) + \psi(y))\mathrm{d}\gamma(x,y),$$

where we have used $\Pi_{X\#}\gamma = \rho$ to simplify the left-hand side. Since $\nu = \Pi_{Y\#}\gamma = T_{\psi\#}\rho$, applying change of variable formulas we get

$$\int_{X\times Y} \psi(y)\mathrm{d}\gamma(x,y) = \int_Y \psi(y)\mathrm{d}\nu = \int_X \psi(T_\psi(x))\rho(x)\mathrm{d}x$$

Substracting this equality from the inequality above shows that the map T_ψ is optimal in the optimal transport :

$$\int_X c(x, T_\psi(x))\rho(x)\mathrm{d}x \leqslant \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) \qquad \square$$

Monge-Ampère equation. Proposition 37 implies that any map T_ψ induced by a Laguerre tessellation of the domain solves the optimal transport between ρ and the image measure $\nu_\psi = T_{\psi\#}\rho$. From now on, we will denote

$$\begin{aligned} G_y &: \mathbb{R}^Y \to \mathbb{R}, \psi \mapsto \rho(\mathrm{Lag}_y(\psi)) \\ G &: \mathbb{R}^Y \to \mathbb{R}^Y, \psi \mapsto (y \mapsto G_y(\psi)). \end{aligned} \tag{4.54}$$

In the bakery analogy, the function $G_y(\psi)$ measures the number of customers for the bakery y given a family of prices $\psi \in \mathbb{R}^Y$, and $G : \mathbb{R}^Y \to \mathbb{R}^Y$ maps a family of prices to a distribution of customers among the bakeries. By (4.53), one has

$$\nu_\psi = \sum_{y\in Y} G_y(\psi)\delta_y.$$

For simplicity, we consider $\mathcal{P}(Y)$ as a subset of $\mathbb{R}^Y$, conflating a probability measure $\nu = \sum_{y\in Y} \nu_y \delta_y$ with the function $\nu : y \mapsto \nu_y$. Then, T_ψ is an optimal transport map between ρ and ν iff $T_{\psi\#}\rho = \nu$ iff

$$G(\psi) = \nu. \tag{4.55}$$

In other words, we have transformed the optimal transport problem into a finite-dimensional non-linear system of equations (4.55).

Remark 18 (Relation to subdifferential and Monge-Ampère equation). Assume that $X = \Omega_X = \mathbb{R}^d$ and that $c(x,y) = -\langle x|y\rangle$. Then,

$$\begin{aligned} \mathrm{Lag}_y(\psi) &= \{x \in \Omega_X \mid \forall z \in Y,\ -\langle x|y\rangle + \psi(y) \leqslant -\langle x|z\rangle + \psi(z)\} \\ &= \{x \in \Omega_X \mid \forall z \in Y,\ \psi(z) \geqslant \langle x|z-y\rangle + \psi(y)\} \end{aligned}$$

Denote $\hat{\psi}$ the convex envelope of ψ, which can be defined using the double Legendre-Fenchel transform by

$$\varphi(x) = \max_{y\in Y} \langle x|y\rangle - \psi(y),$$

$$\hat{\psi}(z) = \max_{x\in X} \langle x|y\rangle - \varphi(x).$$

Then the Laguerre cells defined above agree with the subdifferential of $\hat{\psi}$, i.e. $\mathrm{Lag}_y(\psi) = \partial\hat{\psi}(y)$. Moreover, in the context of Monge-Ampère equations, the (infinite) measure

$$\sum_{y\in Y} \lambda(\mathrm{Lag}_y(\psi))\delta_y = \sum_{y\in Y} \lambda(\partial\hat{\psi}(y))\delta_y$$

is called the *Monge-Ampère measure* of the function $\hat{\psi}$ [57]. Semi-discrete techniques can also be applied to the numerical resolution of Monge-Ampère equations (with e.g. Dirichlet boundary conditions). We refer the reader to the pioneering work of Oliker-Prussner [79] and to the survey by Neilan, Salgado and Zhang [77].

Remark 19 (Lack of uniqueness). The solution ψ to $G(\psi) = \nu$ is never unique, because G is invariant under addition of a constant (see Proposition 38-(iii)). When $\mathrm{spt}(\rho)$ is disconnected there might also exist two solutions ψ^0, ψ^1 to $G(\psi^i) = \nu$ such that $\psi^0 - \psi^1$ is not constant. Take $X = [-1, 1]$, $Y = \{-1, 1\}$, choose $c(x, y) = (x - y)^2$ and

$$\rho = \mathbf{1}_{[-1,-\frac{1}{2}]\cup[\frac{1}{2},1]} \qquad \nu = \frac{1}{2}(\delta_{-1} + \delta_1),$$

A computation shows that if $|\psi(1) - \psi(-1)| \leqslant 2$, then $G(\psi) = \nu$.

The existence of solutions to (4.55) and the algorithms that one can use to solve this system depend crucially on the properties of the function G. In the next proposition, we denote $(\mathbf{1}_y)_{y\in Y}$ the canonical basis of $\mathbb{R}^Y$, i.e. $\mathbf{1}_y(z) = 1$ if $y = z$ and 0 if not. We also denote $\mathbf{1}_Y$ the constant function on Y equal to 1. On $\mathbb{R}^Y$ we consider two norms:

$$\|\psi\| = \sqrt{\sum_{y\in Y} |\psi(y)|^2} \quad \text{and} \quad \|\psi\|_\infty = \max_{y\in Y} |\psi(y)|\,.$$

We will often use the notation R, which measures the oscillation of the cost function:

$$R := \max_{X\times Y} c - \min_{X\times Y} c, \tag{4.56}$$

Proposition 38. *Assume c is twisted (Def. 8) and $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$. Then,*

(i) $\forall y \in Y, \forall t \geqslant 0$, $G_y(\psi + t\mathbf{1}_y) \leqslant G_y(\psi)$,
(ii) $\forall y \neq z \in Y, \forall t \geqslant 0$, $G_y(\psi + t\mathbf{1}_z) \geqslant G_y(\psi)$,
(iii) $\forall \psi \in \mathbb{R}^Y, \forall t \in \mathbb{R}$, $G(\psi + t\mathbf{1}_Y) = G(\psi)$,
(iv) $\forall \psi \in \mathbb{R}^Y, G(\psi) \in \mathcal{P}(Y)$,
(v) *if* $\psi \in \mathbb{R}^Y$ *is such that* $G_{y_0}(\psi) > 0$, *then* $\psi(y_0) \leqslant \min_Y \psi + R$,
(vi) *if* $\psi \in \mathbb{R}^Y$ *is such that* $G_y(\psi) > 0$ *for every* $y \in Y$, *then* $\max_Y \psi - \min_Y \psi \leqslant R$,
(vii) *G is continuous,*

where $R = \max_{X\times Y} c - \min_{X\times Y} c$.

Proof. The properties (i), (ii), (iii) are straightforward consequences of the definition of Laguerre cells. Property (iv) is a consequence of Proposition 37 and of the assumption $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$. To prove (v), take ψ such that $G_{y_0}(\psi) >$

0, implying in particular that the Laguerre cell $\mathrm{Lag}_{y_0}(\psi)$ is non-empty and contains a point $x \in X$. Then, by definition of the cell one has for all $y \in Y \setminus \{y_0\}$, $c(x, y_0) + \psi(y_0) \leqslant c(x, y) + \psi(y)$, thus showing that $\psi(y_0) \leqslant \min_Y \psi + R$. Point (v) is a consequence of Point (vi).

It remains to establish that each of the maps G_y is continuous. For this purpose, we consider a sequence $(\psi_n)_{n\in\mathbb{N}} \in \mathbb{R}^Y$ converging to some $\psi_\infty \in \mathbb{R}^Y$. We first note that as in the proof of Proposition 37, the set

$$S = \{x \in X \mid \exists y \neq z \in Y \text{ s.t. } c(x, y) + \psi(y) = c(x, y) + \psi(z)\}.$$

is Lebesgue-negligible and therefore also ρ-negligible. Defining $\chi = \mathbf{1}_{\mathrm{Lag}_y(\psi)}$ and $\chi_n = \mathbf{1}_{\mathrm{Lag}_y(\psi_n)}$,

$$G_y(\psi_n) = \int \chi_n \mathrm{d}\rho, \text{ and } G(\psi) = \int \chi \mathrm{d}\rho.$$

To prove that $\lim_{n\to+\infty} G_y(\psi_n) = G_y(\psi)$ it suffices to establish that χ_n converges to χ on $X \setminus S$, which is straightforward (because the inequalities defining the set $X \setminus S$ are strict), and to apply Lebesgue's dominated convergence theorem. □

From these properties of G, we can deduce the existence of a solution to the equation $G(\psi) = \nu$. The strategy used to prove this proposition is borrowed from [24] and is also reminiscent of Perron's method to prove existence to Monge-Ampère equations, see e.g. [57].

Corollary 39. *Let $G : \mathbb{R}^Y \to \mathbb{R}^Y$ satisfying (i)– (vi) in Proposition 38 and let $\nu \in \mathcal{P}(Y)$. Then, there exists $\psi \in \mathbb{R}^Y$ such that $G(\psi) = \nu$.*

Proof. Fix some $y_0 \in Y$ such that $\nu_{y_0} \neq 0$, and consider the set

$$K = \{\psi \in \mathbb{R}^Y \mid \psi(y_0) = 0 \text{ and } \forall y \in Y \setminus \{y_0\}, G_{y_0}(\psi) \leqslant \nu_y \text{ and } \psi(y) \leqslant R\},$$

where R is defined as in (4.56). Given $\psi \in K$, one has

$$G_{y_0}(\psi) = 1 - \sum_{y\neq y_0} G_y(\psi) \geqslant \nu_{y_0} > 0,$$

implying by (v) that $\min_{y\in Y} \psi \geqslant \psi(y_0) - R$. The set K is therefore bounded and closed (by continuity of the functions G_y) and therefore compact. We consider ψ^* a minimizer over the set K of the function $J(\psi) = \sum_{y\in Y} \psi_y$. Assume that $G_y(\psi^*) < \nu_y$ for some $y \in Y \setminus \{y_0\}$. Then, by continuity of G_y, there exists some $t > 0$ such that $G_y(\psi^* - t\mathbf{1}_y) < \nu_y$. Then, by property (ii), we have

$$\forall z \neq y, G_z(\psi^* - t\mathbf{1}_y) \leqslant G_z(\psi^*) \leqslant \nu_z,$$

thus showing that $\psi^* - t\mathbf{1}_y \in K$. Since $J(\psi^* - t\mathbf{1}_y) = J(\psi^*) - t < J(\psi^*)$, we get a contradiction. We thus have showed that $\forall y \in Y \setminus \{y_0\}$, $G_y(\psi^*) = \nu_y$, and using (iv) and $\nu \in \mathcal{P}(Y)$ we obtain

$$G_{y_0}(\psi^*) = 1 - \sum_{y\in Y\setminus\{y_0\}} G_y(\psi^*) = 1 - \sum_{y\in Y\setminus\{y_0\}} \nu_y = \nu_{y_0},$$

so that $G(\psi^*) = \nu$ and ψ^* is a solution to (4.55). □

Kantorovich's functional. We now show that Equation (4.55) is the optimality condition of the Kantorovitch functional, and can thus be recast as a smooth unconstrained optimization problem. We recall that

$$(\mathrm{KP}) = \max_{\psi \in \mathbb{R}^Y} \mathcal{K}(\psi),$$

where $\mathcal{K}$ is the Kantorovich functional given by

$$\begin{aligned}\mathcal{K}(\psi) &= \int_X \psi^c \mathrm{d}\mu - \int_Y \psi \mathrm{d}\nu \\ &= \sum_{y\in Y} \int_{\mathrm{Lag}_y(\psi)} (c(x,y) + \psi(y)) \mathrm{d}\rho(x) - \sum_{y \in Y} \psi(y)\nu_y.\end{aligned}$$

Theorem 40 (Aurenhammer, Hoffman, Aronov)**.** *Assume that $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$, that c is twisted (Def. 8), and consider $\mathcal{K}$ defined in* (2.24)*. Then:*

- *$\mathcal{K}$ is concave and $\mathcal{C}^1$-smooth and its gradient is*
$$\nabla \mathcal{K}(\psi) = G(\psi) - \nu \tag{4.57}$$
where G is defined in (4.54)*.*
- *$\forall \psi \in \mathbb{R}^Y, \forall t \in \mathbb{R},\ \mathcal{K}(\psi + t\mathbf{1}_Y) = \mathcal{K}(\psi)$,*
- *$\mathcal{K}$ attains its maximum over $\mathbb{R}^Y$, and $\nabla\mathcal{K}(\psi) = 0$ iff ψ solves* (4.55)*.*

Remark 20. This theorem could be deduced from the computation of directional derivatives of $\mathcal{K}$ given in Corollary 16, however we prefer to give a simple and self-contained proof due to Aurenhammer, Hoffman, Aronov [6].

Proof of Theorem 40. We simultaneously show that the functional is concave and compute its gradient. For any function ψ on Y and any measurable map $T : X \to Y$, one has

$$\min_{y \in Y} (c(x,y) + \psi(y)) \leqslant c(x, T(x)) + \psi(T(x)),$$

which by integration against ρ gives

$$\mathcal{K}(\psi) \leqslant \int_X (c(x,T(x)) + \psi(T(x)))\rho(x)\mathrm{d}x - \sum_{y\in Y} \psi(y)\nu_y. \tag{4.58}$$

Moreover, equality holds when $T = T_\psi$. Taking another function $\psi' \in \mathbb{R}^Y$ and setting $T = T_{\psi'}$ in Equation (4.58) gives

$$\begin{aligned}\mathcal{K}(\psi) &\leqslant \int_X (c(x, T_{\psi'}(x)) + \psi(T_{\psi'}(x)))\rho(x)\mathrm{d}x - \sum_{y\in Y}\psi(y)\nu_y \\ &= \sum_{y\in Y}\int_{\mathrm{Lag}_y(\psi')} (c(x,y) + \psi(y))\rho(x)\mathrm{d}x - \sum_{y\in Y}\psi(y)\nu_y \\ &= \sum_{y\in Y}\int_{\mathrm{Lag}_y(\psi')} (c(x,y) + \psi'(y))\rho(x)\mathrm{d}x + \\ &\quad \sum_{y\in Y} \rho(\mathrm{Lag}_y(\psi'))(\psi(y) - \psi'(y)) - \sum_{y\in Y}\psi(y)\nu_y \\ &= \mathcal{K}(\psi') + \langle G(y) - \nu | \psi - \psi'\rangle\end{aligned}$$

By definition, this shows that $G(\psi) - \nu$ belongs to the superdifferential to $\mathcal{K}$ (Definition 21) at ψ, i.e. $G(\psi) - \nu \in \partial^+\mathcal{K}(\psi)$, thus proving by Proposition 55 that $\mathcal{K}$ is concave.

We now prove that $\mathcal{K}$ belongs to $\mathcal{C}^1(\mathbb{R}^Y)$. Consider $\psi \in \mathbb{R}^Y$ and let $(\psi_n)_{n\in\mathbb{N}}$ be a sequence converging to ψ and such that $\nabla\mathcal{K}(\psi_n)$ exists for every $n \in \mathbb{N}$. Since $G(\psi_n) - \nu \in \partial^+\mathcal{K}(\psi_n) = \{\nabla\mathcal{K}(\psi_n)\}$, we obtain $\nabla\mathcal{K}(\psi_n) = G(\psi_n) - \nu$. Thus, by the continuity of G (Proposition 38),

$$\lim_{n\to+\infty} \nabla\mathcal{K}(\psi_n) = \lim_{n\to+\infty} G(\psi_n) - \nu = G(\psi) - \nu,$$

ensuring by (5.76) that $\partial^+G(\psi) = \{G(\psi) - \nu\}$, so that $\nabla\mathcal{K}(\psi) = G(\psi) - \nu$ for all $\psi \in \mathbb{R}^Y$. By continuity of G we get $\mathcal{K} \in \mathcal{C}^1$ as announced, and Equation (4.54) holds for all $\psi \in \mathbb{R}^Y$, so that one trivially has $\nabla\mathcal{K}(\psi) = 0$ iff $G(\psi) = \nu$. Finally, we note that thanks to Corollary 39, there exists $\psi \in \mathbb{R}^Y$ such that $G(\psi) = \nu$, which automatically is a maximizer of $\mathcal{K}$ because $\mathcal{K}$ is concave and $\nabla\mathcal{K}(\psi) = 0$. $\square$

4.2. **Semi-discrete optimal transport via coordinate decrements.** As before, we assume that $X \subseteq \Omega_X$ is compact, that $Y \subseteq \Omega_Y$ is finite and that $\Omega_X, \Omega_Y \subseteq \mathbb{R}^d$ are open sets. We recall the notation $G_y(\psi) := \rho(\mathrm{Lag}_y(\psi))$. Oliker-Prussner's algorithm for solving $G(\psi) = \nu$ is described in Algorithm 3, and bears strong resemblance with Bertsekas' auction algorithm, in that the "prices" are evolved in a monotonic way.

Algorithm 3 Oliker-Prussner algorithm

Input: A tolerence parameter $\delta > 0$.

Initialization: Fix some $y_0 \in Y$ once for all. Set

$$\psi^{(0)}(y) := \begin{cases} 0 & \text{if } y = y_0 \\ R & \text{if not.} \end{cases}$$

While: $\exists y \neq y_0$ such that $G_y(\psi^{(k)})) \leqslant \nu_y - \frac{\delta}{N}$

Step 1: Compute

$$t_y = \min\{t \geqslant 0 \mid G_y(\psi^{(k)} - t\mathbf{1}_y) \geqslant \nu_y\}. \tag{4.59}$$

Step 2: Set $\psi^{(k+1)} = \psi^{(k)} - t\mathbf{1}_y$.

Output: A vector $\psi^{(k)}$ that satisfies $\left\|G(\psi^{(k)}) - \nu\right\|_\infty \leqslant \delta$.

This algorithm can be described in words using the bakery analogy of Section 4.1. We choose once and for all a bakery $y_0 \in Y$ whose price will be set to zero. Initially, the price of bread $\psi^{(0)}$ is zero at this bakery y_0 and set to the prohibitively large value R, defined in Equation (4.56), at any other location. This choice guarantees that the bakery y_0 initially gets all the customers. The prices $\psi^{(k)} \in \mathbb{R}^y$ are then constructed iteratively by performing a sort of reverse auction: at step k, start by finding some bakery $y = y^{(k)} \in Y \setminus \{y_0\}$ which sells less bread than its production capacity, i.e.

$$G_y(\psi^{(k)}) \leqslant \nu_y - \frac{\delta}{N}.$$

The price of bread at y is then decreased so that the amount of bread sold equals the production capacity of y, i.e. one finds $t_y \geqslant 0$ such that

$$G_y(\psi^{(k)} - t_y \mathbf{1}_y) = \nu_y$$

and then updates $\psi^{(k+1)} = \psi^{(k)} - t_y \mathbf{1}_y$.

Remark 21 (Origin and extensions). This algorithm was introduced by Oliker and Prussner, for the purpose of solving Monge-Ampère equations with Dirichlet boundary conditions in [79]. In the context of optimal transport, the first use of Algorithm 3 seems to be in an article of Caffarelli, Kochengin and Oliker [26] (see also [24]), in the setting of the reflector problem, namely $c(x,y) = -\log(1 - \langle x|y\rangle)$ on $X = Y = \mathcal{S}^{d-1}$. Since then, the convergence of this algorithm has been generalized to more other costs and/or more general assumptions on the probability density ρ, we refer the reader to [64, 42] and to references therein.

$\mathcal{C}^{1,1}$ estimates for Kantorovich functional. The proof of convergence of Oliker-Prussner's algorithm relies on the Lipschitz regularity of the map G when ρ is bounded, proven in the next proposition. (Since $\nabla\mathcal{K} = G - \nu$, this proposition also implies that Kantorovich's functional $\mathcal{K}$ has Lipschitz gradient, improving from the $\mathcal{C}^1$ estimate of Theorem 40.)

Proposition 41. *Assume that $c \in \mathcal{C}^2(\Omega_X \times \Omega_Y)$ satisfies the twist condition, and assume also that $\rho \in \mathcal{P}^{\mathrm{ac}}(X) \cap \mathrm{L}^\infty(X)$. Then for every $y \in Y$, the map $G_y : \mathbb{R}^Y \to \mathbb{R}$ defined in (4.54) is globally Lipschitz.*

Remark 22. The proof of this proposition comes with an estimation of the Lipschitz constant: namely it shows $|G_y(\psi) - G_y(\varphi)| \leqslant L_G \|\varphi - \psi\|_\infty$ with

$$\begin{aligned}
L_G &= c(d) N \|\rho\|_\infty \frac{1}{\kappa}\left(1 + \frac{M}{\kappa}\operatorname{diam}(X)\right)\operatorname{diam}(X)^{d-1}, \\
\kappa &= \min_{y \neq z \in Y} \min_X \|\nabla_x c(\cdot, y) - \nabla_x c(\cdot, z)\|, \\
M &= \max_{y \neq z \in Y} \max_X \left\|\mathrm{D}^2_{xx} c(\cdot, y) - \mathrm{D}^2_{xx} c(\cdot, z)\right\|.
\end{aligned} \tag{4.60}$$

In the estimation of the Lipschitz constant L_G (4.60), it is possible that the term in N is not tight, but the other terms cannot be improved without adding assumptions on the cost.

Example 6. With $c(x,y) = \frac{1}{2}\|x - y\|^2$, one has $\nabla_x c(x,y) = (x - y)$ and $\mathrm{D}^2_{xx} c(x,y) = \mathrm{id}$, so that $M = 0$ and κ is the minimal distance between two distinct points in Y: $\kappa = \min_{y \neq z \in Y} \|y - z\|$.

The proof relies on the following lemma, which allows to estimate the variations of G_y in the direction $\mathbf{1}_z$, $z \neq y$.

Lemma 42. *Let $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$ be a twisted cost and $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$. For every $y \neq z \in Y$ and $\psi \in \mathbb{R}^Y$,*

$$G_y(\psi + t\mathbf{1}_z) - G_y(\psi) = \int_0^t G_{yz}(\psi + s\mathbf{1}_z)\mathrm{d}s. \tag{4.61}$$

where

$$G_{yz}(\psi) = \int_{\mathrm{Lag}_{yz}(\psi)} \frac{\rho(x)}{\|\nabla_x c(y,x) - \nabla_x c(y,z)\|} \mathrm{dvol}^{d-1}(x).$$

Proof. This is a consequence of the coarea formula, Equation (5.78). In order to see this, we first note that

$$\mathrm{Lag}_y(\psi + t\mathbf{1}_z) = E \cap H^{\leqslant}_{yz}(\psi) \quad \text{where } E = \bigcap_{w\in Y\setminus\{y,z\}} H^{\leqslant}_{yw}(\psi).$$

In particular, for $t \geqslant 0$, setting $c_{yz} = c(\cdot, y) - c(\cdot, z)$ and $a = \psi(z) - \psi(y)$,

$$\mathrm{Lag}_y(\psi + t\mathbf{1}_z) \setminus \mathrm{Lag}_y(\psi) = E \cap c_{yz}^{-1}((a, a+t])$$

Thus, by the coarea formula,

$$\begin{aligned} G_y(\psi + t\mathbf{1}_z) - G_y(\psi) &= \rho(\mathrm{Lag}_y(\psi + t\mathbf{1}_z) \setminus \mathrm{Lag}_y(\psi)) \\ &= \int_{E\cap c_{yz}^{-1}((a,a+t])} \rho(x) \mathrm{d}\,\mathrm{vol}^d(x) \\ &= \int_0^t \int_{E\cap c_{yz}^{-1}(a+s)} \frac{\rho(x)}{\|\nabla c_{yz}(x)\|} \mathrm{d}\mathrm{vol}^{d-1}(x)\mathrm{d}s \end{aligned}$$

One concludes by remarking that

$$\begin{aligned} x \in \mathrm{Lag}_{yz}(\psi + s\mathbf{1}_z) &\Longleftrightarrow x \in E \text{ and } c(x,y) + \psi(y) = c(x,z) + \psi(z) + s \\ &\Longleftrightarrow x \in E \cap c_{yz}^{-1}(a+s). \end{aligned}$$

This establishes (4.61) in the case $t \geqslant 0$, and the case $t \leqslant 0$ can be treated similarly. □

The second ingredient to prove Proposition 41 is an uniform upper bound on the $(d-1)$–Hausdorff measure of the level set of a $\mathcal{C}^2$ function f with non-vanishing gradient.

Lemma 43. *Let $X \subseteq \Omega_X \subseteq \mathbb{R}^d$ with Ω_X open and X compact, and let $f \in \mathcal{C}^2(\Omega_X)$ such that $\forall x \in \Omega_X, \|\nabla f(x)\| > 0$. Then,*

$$\mathrm{vol}^{d-1}(f^{-1}(0) \cap X) \leqslant c(d) \left(1 + \frac{M}{\kappa}\,\mathrm{diam}(X)\right) \mathrm{diam}(X)^{d-1}.$$

where $\kappa = \min_X \|\nabla f\|$ and $M = \max_X \left\|\mathrm{D}^2 f\right\|$.

Proof. By compactness, there exists a finite number of unit vectors $u_1, \dots u_n$ and $V_1, \dots, V_n$ an open covering of the unit sphere $\mathcal{S}^{d-1}$ such that if $u \in V_i$, then $\langle u|u_i\rangle \geqslant 3/4$, implying in particular, $\|u - u_i\|^2 = 2 - 2\langle u|u_i\rangle \leqslant \frac{1}{2}$. Moreover n depends only on the dimension d. Let $S = f^{-1}(0) \cap X$. This set S can be covered by patches S_i, i.e. $S = \cup_i S_i$ where

$$S_i = \{x \in S \mid \nabla f(x) \in V_i\}.$$

We will now estimate the volume of each patch S_i using the coarea formula recalled in Theorem 56 of the appendix. To apply this formula, we consider $\Pi_i : S_i \subseteq \mathbb{R}^d \to \{u_i\}^\perp$ the orthogonal projection onto the hyperplane $H_i = \{u_i\}^\perp$. We need to estimate the Jacobian $J_{\Pi_i}(x)$ (see (5.77)). Since Π_i is linear, we have $\mathrm{D}\Pi_i = \Pi_i$. Moreover, for any tangent vector v at $x \in S_i$, one has $\langle v|\nabla f(x)\rangle = 0$. Setting $u = \nabla f(x) \in V_i$, we get

$$\begin{aligned}\|\Pi_i v\|^2 &= \|v\|^2 - \langle v|u_i\rangle^2\\ &= \|v\|^2 - \langle v|u_i - u\rangle^2\\ &\geqslant \|v\|^2 (1 - \|u_i - u\|^2) \geqslant \frac{1}{2}\|v\|^2 .\end{aligned}$$

This directly shows that the restriction of $\mathrm{D}\Pi_i(x)$ to the tangent space $T_x S_i$ at S_i is injective and that its inverse is $\frac{1}{\sqrt{2}}$-Lipschitz. This implies that

$$J_{\Pi_i}(x) \geqslant c(d) = \left(\frac{1}{\sqrt{2}}\right)^{d-1}.$$

We now apply the co-area formula (5.78) to the manifold $M = f^{-1}(0)$, $E = S_i \subseteq N$, $N = H_i$, $n = m = d-1$, $\Phi = \Pi_i$, and $u \equiv 1$:

$$\begin{aligned}\mathrm{vol}^{d-1}(S_i) &= \int_{H_i}\int_{\Pi_i^{-1}(y)} \frac{1}{\mathrm{J}_{\Pi_i}(x)} \mathrm{dvol}^0(x)\mathrm{dvol}^{d-1}(y)\\ &\leqslant c(d)\int_{\{u_i\}^\perp} \mathrm{Card}(S_i \cap (y + \mathbb{R}u_i))\mathrm{dvol}^{d-1}(y).\end{aligned}$$

We now give an upper bound on $\mathrm{Card}(S_i \cap (y + \mathbb{R}u_i))$. Let $x \in S_i$ and use Taylor's formula to get

$$f(x + tu_i) \geqslant f(x) + t\langle \nabla f(x)|u_i\rangle - \frac{M}{2}t^2 \geqslant \frac{3}{4}\kappa t - \frac{M}{2}t^2$$

so that $f(x + tu_i) > 0$ as long as $t \in (0, t^*)$ with $t^* = \frac{3\kappa}{2M}$. One has a similar bound for negative t. This directly implies that the number of intersection points between S_i and $y + \mathbb{R}u_i$ is at most $1 + \mathrm{diam}(X)/t^*$. Since the number n of directions u_i only depends on the dimension d, we have

$$\begin{aligned}\mathrm{vol}^{d-1}(S) &\leqslant \sum_{1\leqslant i\leqslant n} \mathrm{vol}^{d-1}(S_i)\\ &\leqslant c(d)\sum_{1\leqslant n} \mathrm{vol}^{d-1}(H_i \cap \Pi_i(X))\left(1 + \frac{M}{\kappa}\,\mathrm{diam}(X)\right)\\ &\leqslant c(d)\left(1 + \frac{M}{\kappa}\,\mathrm{diam}(X)\right)\mathrm{diam}(X)^{d-1}.\end{aligned}$$ □

Proof of Proposition 41. Let $y \in Y$. Applying Lemma 42, we have

$$\begin{aligned}|G_y(\psi + t\mathbf{1}_y) - G_y(\psi)| &\leqslant \left|\int_0^t \int_{\mathrm{Lag}_{yz}(\psi + s\mathbf{1}_y)} \frac{\rho(x)}{\|\nabla c_{yz}(x)\|}\mathrm{dvol}^{d-1}(x)\mathrm{d}s\right|\\ &\leqslant \frac{\|\rho\|_\infty}{\kappa} \max_{a\leqslant s\leqslant a+t} \mathrm{vol}^{d-1}(c_{yz}^{-1}(s) \cap X)\,|t|\,,\end{aligned}$$

where we used the bound $\|\nabla c_{yz}(x)\| \geqslant \kappa$, which comes from the twist assumption and the inclusion $\mathrm{Lag}_{yz}(\psi + s\mathbf{1}_z) \subseteq c_{yz}^{-1}(a+s)$ with $a = \psi(z) - \psi(y)$, as in the proof of the previous lemma. Applying Lemma 43 to the function $f = c_{yz} - s$, we get a uniform upper bound on the $(d-1)$-volume of the level set $c_{yz}^{-1}(s)$:

$$\mathrm{vol}^{d-1}(c_{yz}^{-1}(s) \cap X) \leqslant c(d) \left(1 + \frac{M}{\kappa} \operatorname{diam}(X)\right) \operatorname{diam}(X)^{d-1},$$

which yields

$$\begin{aligned} &|G_y(\psi + t\mathbf{1}_z) - G_y(\psi)| \leqslant \hat{L}_G\,|t|\,, \\ &\text{with } \hat{L}_G = \frac{c(d)}{\kappa}\left(1 + \frac{M}{\kappa}\operatorname{diam}(X)\right)\operatorname{diam}(X)^{d-1}\ \|\rho\|_\infty. \end{aligned} \tag{4.62}$$

Take $\psi, \tilde{\psi} \in \mathbb{R}^Y$. Order the points in Y, i.e. let $Y = \{y_1, \ldots, y_N\}$ and define recursively

$$\begin{cases} \psi^0 = \psi \\ \psi^{k+1} = \psi^k + (\tilde{\psi}(y_k) - \psi(y_k))\mathbf{1}_{y_k} \end{cases}$$

Then, $\psi^N = \tilde{\psi}$ and for $k \geqslant 1$, ψ^{k+1} and ψ^k differ only by the value at y_k Thus, applying (4.62),

$$\begin{aligned} \left|G_y(\tilde{\psi}) - G_y(\psi)\right| &= \sum_{1\leqslant k\leqslant N} \left|G_y(\psi^{k+1}) - G_y(\psi^k)\right| \\ &\leqslant \sum_{1\leqslant k\leqslant N} \hat{L}_G \left|\psi(y_k) - \tilde{\psi}(y_k)\right| \\ &\leqslant L_G \left\|\psi - \tilde{\psi}\right\|_\infty \text{ with } L_G = N\hat{L}_G \end{aligned}$$

□

Convergence of Oliker-Prussner's algorithm. Now that we have established the Lipschitz continuity of G_y, the convergence of Algorithm 3 follows easily, using arguments similar to those used to establish the convergence of Auction's algorithm.

Theorem 44 (Oliker-Prussner)**.** *Assume that the cost $c \in \mathcal{C}^2(\Omega_X \times \Omega_Y)$ is twisted (Def. 8) and that $\rho \in \mathcal{P}^{\mathrm{ac}}(X) \cap \mathrm{L}^\infty(X)$. Then,*

- *Oliker-Prussner's algorithm converges in a finite number of steps $k \leqslant \mathrm{C}N^3/\delta$, where C is a constant that depends on X, Y, ρ and c.*
- *Furthermore, at step k, one has*

$$\forall 1 \leqslant i \leqslant N, \left|G_i(\psi^{(k)}) - \nu_i\right| \leqslant \delta.$$

Remark 23 (Computational complexity)*.* The computational complexity is actually much higher than the number of steps of the algorithm, since:

- at each iteration, one needs to compute t_y (this could be done using for instance a binary search or more clever techniques).
- each time the map G_y is evaluated, one needs to compute the Laguerre cell $\mathrm{Lag}_y(\psi)$, which, if done naively, requires to compute the intersection of $N - 1$ half-spaces $H_{yz}^{\leqslant}(\psi)$.

Overall, this leads to an upper bound on computational complexity of at least $\mathrm{O}(\frac{N^4}{\delta}\log(N))$, assuming that one can compute $\mathrm{Lag}_y(\psi)$ in time N. To the best of our knowledge, there exists no lower bound on the number of iterations of Algorithm 3, i.e. specific instances of the problem for which one can count the number of iterations.

Remark 24 (δ-Scaling). It is tempting to perform δ-scaling as in the case of Auction's algorithm (see Algorithm 2). In practice, one could start with a rather large $\delta^{(0)} \in (0,1)$, to get a first estimation of the prices using Oliker-Prussner's algorithm. Then one would iteratively replace $\delta^{(\ell)}$ by $\delta^{(\ell+1)} = \frac{1}{2}\delta^{(\ell)}$ and run again the algorithm starting from the prices found at the previous iteration. Doing so, one could hope to get rid of the $\frac{1}{\delta}$ term in the number of iterations, and to replace it by e.g. $\log\left(\frac{1}{\delta}\right)$.

Proof of Theorem 44.
Step 1 *(Correctness)* When Algorithm 3 terminates with $\psi := \psi^{(k)}$, one has for any $y \neq y_0$, $\rho(\mathrm{Lag}_y(\psi)) \leqslant \nu_y$. When it stops, it also means that one has $\rho(\mathrm{Lag}_y(\psi)) \geqslant \nu_y - \frac{\delta}{N}$. Then, as desired, we get

$$\rho(\mathrm{Lag}_{y_0}(\psi)) = 1 - \sum_{y \neq y_0} \rho(\mathrm{Lag}_{y_0}(\psi)) \in [\nu_{y_0}, \nu_{y_0} + \delta].$$

Step 2 *(A priori bound on ψ_k)* By construction one has $\rho(\mathrm{Lag}_y(\psi^{(k)})) \leqslant \nu_y$, which also imply that

$$\rho(\mathrm{Lag}_{y_0}(\psi^{(k)})) = 1 - \sum_{y \in Y \setminus \{y_0\}} \rho(\mathrm{Lag}_y(\psi^{(k)})) \geqslant \nu_{y_0} > 0.$$

By Proposition 38–(v), we get $0 = \psi^k(y_0) \leqslant \min_Y \psi^{(k)} + R$. Since the price of y_0 is never changed, $\psi^{(k)}(y_0) = 0$ and $R \geqslant \psi^{(k)} \geqslant -R$.
Step 3 *(Minimum decrease and termination)* In the second step of the algorithm, when $\psi^{(k)}$ is updated one has $G_y(\psi^{(k)} - t_t \mathbf{1}_y) \geqslant G_y(\psi^{(k)}) + \frac{\delta}{N}$. Since G_y is Lipschitz with some constant L_G, this implies that $|t_y| \geqslant \frac{\delta}{NL_G}$. Then, since $\psi_0(y) = R$ and for any k, $\psi_k(y) \geqslant -R$, the number of times k_y the price of a point $y \in Y$ has been updated cannot be too large:

$$k_y \delta/(NL_G) \leqslant 2R,$$

i.e. $k_y \leqslant (2RNL_G)/\delta$. Since this bound on the number of steps is for a single point, it needs to be multiplied by N to get the total number of steps. Using the bound on L_G given in (4.60), we get an upper bound of $\mathrm{O}(\frac{N^3}{\delta})$ on the number of iterations of the algorithm. $\square$

4.3. **Semi-discrete optimal transport via Newton's method.** We consider a simple damped Newton's algorithm to solve semi-discrete optimal transport problem introduced in [65], and adapted from a similar algorithm for solving Monge-Ampère equations with Dirichlet boundary conditions [75].

Hessian of Kantorovich's functional. In order to write the Newton's algorithm, we first show that G is $\mathcal{C}^1$ (or equivalently $\mathcal{K}$ is $\mathcal{C}^2$) and we compute its derivatives under a genericity assumption, which depends on the cost and on the choice of points Y. This condition is a bit technical, but is for instance satisfied for the quadratic cost on $\mathbb{R}^d$ (see Remark 27 below).

Definition 18 (Genericity assumption)**.** Let $\Omega_X, \Omega_Y \subseteq \mathbb{R}^d$ open, $c \in \mathcal{C}^1(\Omega_X \times \Omega_Y)$, and $X \subseteq \Omega_X$, $Y \subseteq \Omega_Y$, with X compact and Y finite.

- We call Y *generic with respect to* c if for all distinct $y_0, y_2, y_2 \in Y$ and all $t \in \mathbb{R}^2$, one has

$$\mathrm{vol}^{d-1}\left(\{x \in \Omega_X \mid (c(x,y_1) - c(x,y_0), c(x,y_2) - c(x,y_0)) = t\}\right) = 0.$$

- We call Y *generic with respect to* ∂X if for all distinct $y_0, y_1 \in Y$ and all $t \in \mathbb{R}$, one has

$$\mathrm{vol}^{d-1}\left(\{x \in \Omega_X \mid c(x,y_1) - c(x,y_0) = t\} \cap \partial X\right) = 0.$$

Remark 25 (d=1)*.* The genericity assumption is *never* satisfied in dimension $d = 1$, because it requires that the intersection of the 0-dimensional sets $\{c(\cdot, y_i) - c(\cdot, y_0) = t_i\}$, with $i = 1, 2$ is empty for all $t \in \mathbb{R}^2$. Nonetheless, quasi-Newton methods seem to be quite efficient in this case as well [40].

Remark 26 (Sufficient genericity condition)*.* Assume for all distinct points $y_0, y_1, y_2 \in Y$ and for every $x \in X$, the vectors $\nabla_x c(x,y_1) - \nabla_x c(x,y_0)$ and $\nabla_x c(x,y_2) - \nabla_x c(x,y_0)$ are independent. Then, the implicit function theorem guarantees that for every $t \in \mathbb{R}^2$ the set

$$(c(\cdot,y_1) - c(\cdot,y_0), c(\cdot,y_2) - c(\cdot,y_0))^{-1}(t)$$

is a $(d-2)$ dimensional submanifold, and therefore has zero $(d-1)$–volume. In particular, the set Y is generic with respect to c (but not necessarily with respect to ∂X).

Remark 27 (Quadratic cost)*.* For the quadratic cost $c(x,y) = \frac{1}{2}\|x-y\|^2$, we have $\nabla_x c(x,y_i) - \nabla_x c(x,y_0) = y_0 - y_i$. Using the previous remark, we see that the set Y is generic with respect to c if it does not include three aligned points.

Genericity with respect to the boundary ∂X requires more assumptions. For instance, if X is a strictly convex set (or more generally if the Gaussian curvature is nonzero at any point on ∂X) and if the cost is quadratic, then Y is automatically generic with respect to ∂X. As a second example, we assume that X is a compact convex polyhedron, e.g.

$$X = \{x \in \mathbb{R}^d \mid \forall 1 \leqslant j \leqslant M, \langle x|w_j\rangle \leqslant 1\},$$

where $w_1, \ldots, w_M \in \mathbb{R}^d$. Then Y is generic with respect to ∂X if for all distinct $y_0, y_1 \in Y$ and any $1 \leqslant i \leqslant M$, the vectors $y_1 - y_0$ and w_j are independent.

Example 7 (Non-differentiability of G)*.* When the set Y isn't generic, the map G might be non-differentiable. Consider for instance $Y = \{y_{-1}, y_0, y_1\} \subseteq \mathbb{R}^2$ with $y_i = (i, 0)$, $X = [-1,1]^2$ and $\rho = \frac{1}{4}\,\mathrm{vol}^2\big|_X$. Define a one-parameter family of prices $\psi_t : Y \to \mathbb{R}$ by $\psi_t(y_0) = t$ and $\psi_t(y_{\pm 1}) = 0$. Then, for $t \geqslant 0$,

$$\mathrm{Lag}_{y_0}(\psi_t) = \{x = (x_1, x_2) \in \mathbb{R}^2 \mid |x_1| \leqslant \frac{1}{2}|1-t|\}.$$

$$G_{y_0}(\psi_t) = \rho(\mathrm{Lag}_{y_0}(\psi_t)) = \begin{cases} \frac{1}{2}|1-t| & \text{if } t \leqslant 1 \\ 0 & \text{if not.} \end{cases},$$

showing that the function G_{y_0} is non-differentiable at $t = 1$.

Theorem 45. *If $c \in \mathcal{C}^2(\Omega_X \times \Omega_Y)$ satisfies the twist condition (Def. 8), Y is generic with respect to c and ∂X (Def. 18), and the restriction $\rho_{|X}$ of ρ to X is continuous ($\rho_{|X} \in \mathcal{C}^0(X)$), then the map $G : \mathbb{R}^Y \to \mathbb{R}^Y$ is $\mathcal{C}^1$, and*

$$\begin{aligned} &\forall z \neq y,\ \frac{\partial G_y}{\partial \mathbf{1}_z}(\psi) = G_{yz}(\psi) := \int_{\mathrm{Lag}_{yz}(\psi)} \frac{\rho(x)}{\|\nabla_x c(x,y) - \nabla_x c(x,z)\|} \mathrm{d}x, \\ &\forall y \in Y,\ \frac{\partial G_y}{\partial \mathbf{1}_y}(\psi) = G_{yy}(\psi) := -\sum_{z \in Y\setminus\{y\}} G_{yz}(\psi) \end{aligned} \tag{4.63}$$

where we denote $\mathrm{Lag}_{yz}(\psi) = \mathrm{Lag}_y(\psi) \cap \mathrm{Lag}_z(\psi)$ for $y \neq z$.

The formula that one should expect for the partial derivative of G_y with respect to $\mathbf{1}_z$ ($z \neq y$) is already quite clear from Lemma 42. The main difficulty in order to establish Theorem 45 is to prove that the function G_{yz} defined in (4.63) is continuous.

Lemma 46. *Assume that Y is generic with respect to c and ∂X. Then, for any $y \neq z \in Y$, the function G_{yz} defined in (4.63) is continuous.*

Proof. Let $f = c_{yz} = c(\cdot, y) - c(\cdot, z)$. By Cauchy-Lipschitz's theory, one can construct a flow $\Phi : [-\varepsilon, \varepsilon] \times \Omega_X^1 \to \Omega_X$, where $\varepsilon > 0$, such that

$$\begin{cases} \Phi(0,x) = x \\ \dot{\Phi}(t,x) = \frac{\nabla f(\Phi(t,x))}{\|\nabla f(\Phi(t,x))\|^2}, \end{cases} \tag{4.64}$$

where $\dot{\Phi}$ is the derivative with respect to t and $\Omega_X^1 \subset \Omega_X$ is an open set containing X. A simple calculation shows that $\frac{\mathrm{d}}{\mathrm{d}t} f(\Phi(t,x)) = 1$, which implies that $f(\Phi(t,x)) = f(\Phi(0,x)) + t$. Moreover, since $\nabla f / \|\nabla f\|$ is of class C^1 on Ω_X, then $F_t := \Phi(t, \cdot)$ converges pointwise in a $\mathcal{C}^1$ sense to the identity as $t \to 0$.

Let (ψ_n) be a sequence in $\mathbb{R}^Y$ converging to some $\psi_\infty \in \mathbb{R}^Y$. We put $a_n = \psi_n(z) - \psi_n(y)$, $a = \psi_\infty(z) - \psi_\infty(y)$ and $t_n = a_n - a$ and define

$$L_n = \Phi(-t_n, \mathrm{Lag}_{yz}(\psi_n)) \quad \text{and} \quad L_\infty = \mathrm{Lag}_{yz}(\psi_\infty).$$

By definition, one has $f(\mathrm{Lag}_{yz}(\psi_n)) = a_n$ and $f(\mathrm{Lag}_{yz}(\psi_\infty)) = a$. Using the flow property, one gets that both L_n and L_∞ are subsets of the hypersurface $H = f^{-1}(a)$. Denoting F_n the restriction of $\Phi(t_n, \cdot)$ to H, one has $\mathrm{Lag}_{yz}(\psi_n) = F_n(L_n)$.

We now need to consider a continuous extension $\overline{\rho}$ of $\rho_{|X}$ onto Ω_X, since L_n may not be included in X. By a change of variable (see (5.79) for instance), one gets

$$\begin{aligned} g_n := G_{yz}(\psi_n) &= \int_{\mathrm{Lag}_{yz}(\psi_n)} \frac{\rho(y)}{\|\nabla c_{yz}(y)\|} \mathrm{dvol}^{d-1}(y) \\ &= \int_{\mathrm{Lag}_{yz}(\psi_n)} \frac{\overline{\rho}(y)}{\|\nabla c_{yz}(y)\|} \chi_X(y) \mathrm{dvol}^{d-1}(y) \\ &= \int_H \frac{\overline{\rho}(F_n(x))}{\|\nabla c_{yz}(F_n(x))\|} JF_n(x) \chi_{L_n}(x) \chi_X(F_n(x)) \mathrm{dvol}^{d-1}(x), \end{aligned}$$

where χ_A is the indicator function of A. Moreover,

$$g_\infty := G_{yz}(\psi) = \int_H \frac{\rho(x)}{\|\nabla c_{yz}(x)\|}\chi_{L_\infty\cap X}(x)\mathrm{dvol}^{d-1}(x).$$

By Lebesgue's dominated convergence theorem, to prove that $(g_n)_{n\geqslant 0}$ converges to g_∞, it suffices to prove that the integrand of g_n (seen as a function on H) tends to the integrand of g_∞ vol^{d-1}-almost everywhere. Since F_n converges to the identity in a C^1 sense and $\overline{\rho}$ is continuous, it remains to show that $\lim_{n\to\infty}\chi_{L_n}(x)\chi_X(F_n(x)) = \chi_{L_\infty\cap X}(x)$ for almost every $x\in H$ (for the $(d-1)$ Hausdorff measure).

We first prove that $\limsup_{n\to\infty}\chi_{L_n}(x))\chi_X(F_n(x)) \leqslant \chi_{L_\infty\cap X}(x)$ for every $x\in H$. The limsup is non-zero if and only if there exists a subsequence $\sigma(n)$ such that $x\in L_{\sigma(n)}$ and $F_n(x)\in X$. Then, since $F_{\sigma(n)}(L_{\sigma(n)}) = \mathrm{Lag}_{yz}(\psi_{\sigma(n)})$ we get

$$\begin{cases} c(F_{\sigma(n)}(x),y)+\psi_{\sigma(n)}(y) \leqslant c(F_{\sigma(n)}(x),w)+\psi_{\sigma(n)}(w)\\ c(F_{\sigma(n)}(x),y)+\psi_{\sigma(n)}(y) = c(F_{\sigma(n)}(x),z)+\psi_{\sigma(n)}(z).\end{cases}$$

Passing to the limit $n\to+\infty$, we see that x belongs to $\mathrm{Lag}_{yz}(\psi_\infty) = L_\infty$ and to X, thus ensuring

$$\limsup_{n\to+\infty}\chi_{L_n}(x)\chi_X(F_n(x)) \leqslant \chi_{L_\infty\cap X}(x).$$

We now pass to the liminf inequality. Denote

$$S = \left(\bigcup_{w\in Y\setminus\{y,z\}} H_{yzw}(\psi_\infty)\right)\cup\left(H_{yz}(\psi)\cap\partial X\right),$$

where H_{yz} is defined in Equation (4.51) and $H_{yzw}(\psi_\infty) := H_{yz}(\psi_\infty)\cap H_{zw}(\psi_\infty)$ which by assumption has zero $(d-1)$ Hausdorff measure. We now prove that $\liminf_{n\to\infty}(x)\chi_{L_n}\chi_X(F_n(x)) \geqslant \chi_{L_\infty\cap X}$ on $H\setminus S$. If $x\notin L_\infty\cap X$, $\chi_{L_\infty\cap X}(x) = 0$ and there is nothing to prove. We therefore consider $x\in(L_\infty\cap X)\setminus S$, meaning by definition of S that x belongs to the interior $\mathrm{int}(X)$ and that

$$\forall w\in Y\setminus\{z,y\}, c(x,y)+\psi_\infty(y) < c(x,w)+\psi_\infty(w).$$

Since $F_n(x)$ converges to x, this implies that for n large enough one has $F_n(x)\in\mathrm{int}(X)$ and

$$\forall w\in Y\setminus\{z,y\}, c(F_n(x),y)+\psi_\infty(y) < c(F_n(x),w)+\psi_\infty(w).$$

By definition, this means that $F_n(x)$ belongs to $\mathrm{Lag}_{yz}(\psi_n)$, and therefore $x\in L_n$ by definition of L_n. Thus

$$\liminf_{n\to+\infty}\chi_{L_n}(x)\chi_X(F_n(x)) = 1 \geqslant \chi_{L_\infty\cap X}(x)$$

$\square$

Proof of Theorem 45. Lemma 42 shows that for any distinct point $y\neq z\in Y$ and any $\psi\in\mathbb{R}^Y$ one has

$$G_y(\psi+t\mathbf{1}_z) = G_y(\psi) + \int_0^t G_{yz}(\psi+s\mathbf{1}_z)\mathrm{d}s.$$

Moreover, by Lemma 46, we know that the function G_{yz} is continuous. The fundamental theorem of calculus implies that $f : t \mapsto G_y(\psi + t\mathbf{1}_z)$ is differentiable, and that $f'(0) = \frac{\partial G_y}{\partial \mathbf{1}_z}(\psi) = G_{yz}(\psi)$. To compute the partial derivative of G_y with respect to $\mathbf{1}_y$, we note that by invariance of G_y under addition of a constant,

$$G_y(\psi + t\mathbf{1}_y) = G_y(\psi - \sum_{z \neq y} t\mathbf{1}_z)$$

The right-hand side of this expression is differentiable with respect to t, so that the left-hand side is also differentiable, and the chain rule gives

$$\frac{\partial G_y}{\partial \mathbf{1}_y}(\psi) = - \sum_{z \in Y \setminus \{y\}} G_{yz}(\psi).$$

Using again the continuity of G_{yz} on $\mathbb{R}^Y$, we obtain $G \in \mathcal{C}^1(\mathbb{R}^Y)$. □

Strong concavity of Kantorovich's functional. We show here a strict monotonicity property of G, which corresponds to a concavity property on the Kantorovitch functional $\mathcal{K}$, since we have $D^2\mathcal{K} = DG$.

Theorem 47. *Assume that c, X, Y, ρ are as in Theorem 45, and in addition that $\rho(\partial X) = 0$ and that the set $\{\rho > 0\} \cap \mathrm{int}(X)$ is connected. Define*

$$\mathcal{S}_+ := \{\psi \in \mathbb{R}^d \mid \forall y \in Y, G_y(\psi) > 0\}.$$

$$\mathcal{S}_\epsilon := \{\psi \in \mathbb{R}^d \mid \forall y \in Y, G_y(\psi) \geqslant \epsilon\}.$$

- *Kantorovich's functional is locally strongly concave on $\mathcal{S}_+ \cap \{\mathbf{1}_Y\}^\perp$:*

$$\forall \psi \in \mathcal{S}_+,\ \forall v \in \{\mathbf{1}_Y\}^\perp \setminus \{0\},\ \langle DG(\psi)v | v\rangle < 0$$

- *For every $\epsilon > 0$, the set of functions $\mathcal{S}_\epsilon \cap \{\mathbf{1}_Y\}^\perp$ is compact.*

Definition 19 (Irreducible matrix)**.** A square matrix H is called *irreducible* if and only if the graph induced by H is connected[3], i.e.

$$\forall (a,b) \in \{1, \dots, N\}, \exists i_1 = a, \dots, i_k = b \text{ s.t. } \forall j \in \{1, \dots, k-1\}, H_{i_j, i_{j+1}} \neq 0.$$

Lemma 48. *Let H be a symmetric irreducible matrix such that $H_{ij} \geqslant 0$ if $i \neq j$ and $H_{ii} = -\sum_{j \neq i} H_{ij}$. Then, H is non-positive and $\ker H = \mathbb{R}(1, \dots, 1) = \mathbb{R}\mathbf{1}_Y$.*

Proof. The non-positivity follows from Gershgorin's circle theorem. The lemma will be established if we prove that any vector in the kernel of H is constant. Consider $v \in \ker H$ and let i_0 be an index where v attains its maximum, i.e. $i_0 \in \arg\max_{1 \leqslant i \leqslant n} v_i$. Then using $Hv = 0$, and in particular $(Hv)_{i_0} = 0$, one has

$$0 = \sum_{i \neq i_0} H_{i,i_0} v_i + H_{i_0,i_0} v_{i_0} = \sum_{i \neq i_0} H_{i,i_0} v_i - \sum_{i \neq i_0} H_{i,i_0} v_{i_0} = \sum_{i \neq i_0} H_{i,i_0}(v_i - v_{i_0}).$$

This follows from $H_{i_0,i_0} = -\sum_{i \neq i_0} H_{i,i_0}$. Since for every $i \neq i_0$, one has $H_{i,i_0} \geqslant 0$ and $v_{i_0} - v_i \geqslant 0$, this implies that $v_i = v_{i_0}$ for every i such that $H_{i,i_0} \neq 0$. By induction and using the connectedness of the graph induced by H, this shows that v has to be constant. □

[3]the graph induced by the $N \times N$ matrix H is the graph with vertices $\{1, \dots, N\}$, and where i, j are linked by an edge if $H_{ij} \neq 0$

Lemma 49. *Let $U \subseteq \mathbb{R}^d$ be a connected open set, and $S \subseteq \mathbb{R}^d$ be a closed set such that* $\mathrm{vol}^{d-1}(S) = 0$. *Then, $U \setminus S$ is path-connected.*

Proof. It suffices to treat the case where U is an open ball, the general case will follow by standard connectedness arguments. Let $x, y \in U \setminus S$ be distinct points. Since $U \setminus S$ is open, there exists $r > 0$ such that $\mathrm{B}(x, r)$ and $B(y, r)$ are included in $U \setminus S$. Consider H the hyperplane orthogonal to the segment $[x, y]$, and Π_H the projection on H. Then, since Π_H is 1-Lipschitz, $\mathrm{vol}^{d-1}(\Pi_H S) \leqslant \mathrm{vol}^{d-1}(S) = 0$, so that $H \setminus \Pi_H S$ is dense in the hyperplane H. In particular, there exists a point

$$z \in \Pi_H(B(x, r)) \setminus S = \Pi_H(B(y, r)) \setminus S.$$

By construction the line $z + \mathbb{R}(y - x)$ avoids S and passes through the balls $\mathrm{B}(x, r) \subseteq U \setminus S$ and $\mathrm{B}(y, r) \subseteq U \setminus S$. This shows that the points x, y can be connected in $U \setminus S$. □

Proof of Theorem 47. Let $Y = \{y_1, \dots, y_N\}$. Fix $\psi \in \mathcal{S}_+$ and define

$$H_{ij} := \frac{\partial G_{y_i}}{\partial \mathbf{1}_{y_j}}(\psi).$$

By Lemma 48, the first claim will hold if we prove that the matrix H is irreducible. We define $Z = \mathrm{int}(X) \cap \{\rho > 0\}$, which by assumption is a connected open set.

Step 1: We show here that for all $i \in \{1, \dots, N\}$, $\mathrm{int}(\mathrm{Lag}_{y_i}(\psi)) \cap Z$ contains at least a point which we denote x_i. Indeed, since $\psi \in \mathcal{S}_+$, we know that $\rho(\mathrm{Lag}_{y_i}(\psi)) > 0$. In addition, by Proposition 37, $\rho(\mathrm{Lag}_{y_i}(\psi) \cap \mathrm{Lag}_{y_j}(\psi)) = 0$ for all $j \neq i$, and $\rho(\partial X) = 0$ by assumption. This implies that $\rho(L_i) = \rho(\mathrm{Lag}_{y_i}(\psi)) > 0$, where

$$L_i = \{x \in Z \mid \forall j \neq i, c(x, y_i) + \psi(y_i) < c(x, y_j) + \psi(y_j)\} \subseteq \mathrm{int}(\mathrm{Lag}_{y_i}(\psi)).$$

We conclude by remarking that L_i is contained in $\mathrm{int}(\mathrm{Lag}_{y_i}(\psi)) \cap Z$, which therefore has to be nonempty.

Step 2: Let S the union of facets that are common to at least three distinct Laguerre cells, i.e.

$$S = \bigcup_{y_1, y_2, y_3 \mathrm{distinct}} \mathrm{Lag}_{y_1, y_2, y_3}(\psi).$$

Then, $Z \setminus S$ is open and path-connected. Indeed, by the genericity assumption (Def 18), we already know that $\mathrm{vol}^{d-1}(S) = 0$, and Lemma 49 then implies that $Z \setminus S$ is path-connected.

Step 3: Let $x \in Z \setminus S$ be such that $x \in \mathrm{Lag}_{y_i}(\psi) \cap \mathrm{Lag}_{y_j}(\psi)$ for $i \neq j$. Then, $H_{ij} > 0$. To see this, we note that since x belongs to the complement of S,

$$\begin{cases} c(x, y_i) + \psi(y_i) = c(x, y_j) + \psi(y_j), \\ \forall k \notin \{i, j\}, c(x, y_i) + \psi(y_i) < c(x, y_k) + \psi(y_k). \end{cases}$$

This implies that there exists a ball with radius $r > 0$ around x such that

$$\forall x' \in \mathrm{B}(x, r), \forall k \notin \{i, j\}, c(x, y_i) + \psi(y_i) < c(x, y_k) + \psi(y_k),$$

directly implying that

$$H_{yy'}(\psi) \cap \mathrm{B}(x, r) \subseteq \mathrm{Lag}_{y_i y_j}(\psi).$$

By the twist hypothesis and the inverse function theorem, $H_{yy'}(\psi)$ is a $(d-1)$-dimensional submanifold. In addition, $\rho(x) > 0$ because x belongs to Z. This implies that

$$\begin{aligned} H_{ij} &= \int_{\mathrm{Lag}_{y_i y_j}(\psi)} \frac{\rho(x')}{\|\nabla_x c(x', y_i) - \nabla_x c(x', y_j)\|} \mathrm{dvol}^{d-1}(x') \\ &\geqslant \int_{H_{y_i y_j}(\psi) \cap \mathrm{B}(x,r)} \frac{\rho(x')}{\|\nabla_x c(x', y_i) - \nabla_x c(x', y_j)\|} \mathrm{dvol}^{d-1}(x') > 0. \end{aligned}$$

Step 4: We now fix $i \neq j \in \{1, \dots, N\}$ and the points x_i, x_j whose existence is established in Step 1:

$$x_i \in \mathrm{int}(\mathrm{Lag}_{y_i}(\psi)) \cap Z, \quad x_j \in \mathrm{int}(\mathrm{Lag}_{y_j}(\psi)) \cap Z,$$

so that in particular x_i, x_j belongs to $Z \setminus S$. By Step 2, we get the existence of a continuous path $\gamma \in \mathcal{C}^0([0,1], Z \setminus S)$ such that $\gamma(0) = x_i$ and $\gamma(1) = x_j$. We define a sequence $i_k \in \{1, \dots, N\}$ of indices by induction, starting from $i_0 = i$. For $k \geqslant 0$ we define $t_k = \max\{t \in [0,1] \mid \gamma(t) \in \mathrm{Lag}_{y_{i_k}}\}$. If $t_k = 1$ we are done. If not, $\gamma(t_k)$ belongs to exactly two distinct Laguerre cells, and we define $i_{k+1} \neq i_k$ so that $\gamma(t_k) \in \mathrm{Lag}_{y_{i_k}}(\psi) \cap \mathrm{Lag}_{y_{i_{k+1}}}(\psi)$. By definition of t_i as a maximum, the points $y_1, \dots, y_k$ must be distinct, so that $t_\ell = 1$ after a finite number of iterations and then $i_\ell = j$. By Step 3, we get that $H_{y_{i_k} y_{i_{k+1}}} > 0$ for any $k \in \{0, \ell-1\}$, proving that the matrix H is irreducible, thus $\ker \mathrm{D}G(\psi) = \mathbb{R}\mathbf{1}_Y$ by Lemma 48, implying the strict concavity property.

Compactness of $\mathcal{S}_\epsilon \cap \{\mathbf{1}_Y\}^\perp$: By continuity of the function G, this set is closed. By Proposition 38-(vi), $\max_Y \psi - \min_Y$ is bounded on the set $\mathcal{S}_\epsilon$. This implies that $\mathcal{S}_\epsilon \cap \{\mathbf{1}_Y\}^\perp$ is bounded since every function of $\{\mathbf{1}_Y\}^\perp$ has a mean value equal to zero, and is thus compact.

□

Damped Newton algorithm and its convergence.

Proposition 50. *Let $G = (G_1, \dots, G_N) \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R}^N)$ be a function satisfying the following properties:*

1. *(Invariance and image) G is invariant under the addition of a constant, $G_i(\psi) \geqslant 0$ and $\sum_i G_i(\psi) = 1$ for all $\psi \in \mathbb{R}^N$.*
2. *(Compactness) For any $\varepsilon > 0$ the set $\mathcal{S}_\varepsilon \cap \{\mathbf{1}\}^\perp$ is compact, where*

$$\begin{aligned} &\mathcal{S}_\varepsilon := \left\{\psi \in \mathbb{R}^N \mid \forall i,\ G_i(\psi) \geqslant \epsilon\right\} \\ &\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^N \end{aligned}$$

3. *(Strict monotonicity) The matrix $\mathrm{D}G(\psi)$ is symmetric nonpositive, and*

$$\forall \psi \in \mathcal{S}_\varepsilon,\ \forall v \in \{\mathbf{1}\}^\perp \setminus \{0\},\ \langle DG(\psi) v | v\rangle < 0.$$

Then Algorithm 4 terminates in a finite number of steps. More precisely, the iterates $(\psi^{(k)})$ of Algorithm 4 satisfy, for some $\tau^* > 0$,

$$\left\| G(\psi^{k+1}) - \nu \right\| \leqslant \left(1 - \frac{\tau^\star}{2}\right) \left\| G(\psi^k) - \nu \right\|.$$

Algorithm 4 Damped Newton algorithm

Input: A tolerance $\eta > 0$ and an initial $\psi^{(0)} \in \mathbb{R}^Y$ such that

$$\varepsilon := \frac{1}{2} \min \left[\min_{y \in Y} G_y(\psi^{(0)}),\ \min_{y \in Y} \nu_y \right] > 0. \tag{4.65}$$

While: $\left\| G(\psi^{(k)}) - \nu \right\|_\infty \geqslant \eta$

Step 1: Compute $v^{(k)}$ satisfying

$$\begin{cases} \mathrm{D}G(\psi^{(k)}) v^{(k)} = \nu - G(\psi^{(k)}) \\ \sum_{y \in Y} v^{(k)}(y) = 0 \end{cases}$$

Step 2: Determine the minimum $\ell \in \mathbb{N}$ such that $\psi^{(k,\ell)} := \psi^{(k)} + 2^{-\ell} v^{(k)}$ satisfies

$$\begin{cases} \forall y \in Y, G_y(\psi^{(k,\ell)}) \geqslant \varepsilon \\ \left\| G(\psi^{(k,\ell)}) - \nu \right\| \leqslant (1 - 2^{-(\ell+1)}) \left\| G(\psi^{(k)}) - \nu \right\| \end{cases}$$

Step 3: Set $\psi^{(k+1)} = \psi^{(k)} + 2^{-\ell} v^{(k)}$ and $k \leftarrow k + 1$.

Output: A vector $\psi^{(k)}$ that satisfies $\left\| G(\psi^{(k)}) - \nu \right\|_\infty \leqslant \eta$.

Proof.

Estimates. Let $\nu \in \mathbb{R}^N$ be such that $\sum_i \nu_i = 1$. We assume that $\psi^{(0)} \in \mathbb{R}^N \cap \{\mathbf{1}\}^\perp$ is chosen so that

$$\epsilon = \frac{1}{2} \min \left(\min_i G_i(\psi^0), \min_i \nu_i \right) > 0,$$

and we let $\mathcal{S} := \mathcal{S}_\varepsilon \cap \{\mathbf{1}\}^\perp$. Let $\psi \in \mathcal{S}$. By Theorem 47, the matrix $\mathrm{D}G(\psi)$ is symmetric non-positive, and its kernel is the one-dimensional space $\mathbb{R}\mathbf{1}$. Thus, the equation

$$\begin{cases} \mathrm{D}G(\psi) v = \nu - G(\psi), \\ \sum_i v_i = 0, \end{cases}$$

has a unique solution, which we denote $v(\psi)$, and we let $\psi_\tau = \psi + \tau v(\psi)$. By continuity of $\mathrm{D}G(\psi)$ over the compact domain $\mathcal{S}$, the non-zero eigenvalues of $-\mathrm{D}G(\psi)$ lie in $[a, A]$ for some $0 < a \leqslant A < +\infty$. In particular, there exists a constant $M > 0$ such that for all $\psi \in \mathcal{S}$

$$\frac{\|G(\psi) - \nu\|}{A} \leqslant \|v(\psi)\| \leqslant \frac{\|G(\psi) - \nu\|}{a} \leqslant M. \tag{4.66}$$

In particular, the function $F : (\psi, \tau) \in \mathcal{S} \times [0, 1] \mapsto \psi_\tau$ is continuous. Since $\mathcal{S} \times [0, 1]$ is compact, $K := F(\mathcal{S} \times [0, 1])$ is also compact. Then, by uniform continuity of $\mathrm{D}G$ over K, we see that there exists an increasing function ω such that $\lim_{t \to 0} \omega(t) = 0$ and $\|\mathrm{D}G(\psi) - \mathrm{D}G(\psi')\| \leqslant \omega(\|\psi - \psi'\|)$ for all $\psi, \psi' \in K$. Since G is of class $\mathcal{C}^1$, a Taylor expansion in τ gives

$$G(\psi_\tau) = G(\psi + \tau v(\psi)) = (1 - \tau) G(\psi) + \tau \nu + R(\tau) \tag{4.67}$$

where $R(\tau) = \int_0^\tau (\mathrm{D}G(\psi_t) - \mathrm{D}G(\psi))v(\psi)\mathrm{d}t$ is the integral remainder. Then, we can bound the norm of $R(\tau)$ for $\tau \in [0,1]$:

$$\begin{aligned}
\|R(\tau)\| &= \left\| \int_0^\tau (\mathrm{D}G(\psi_t) - \mathrm{D}G(\psi))v(\psi)\mathrm{d}t \right\| \\
&\leqslant \|v(\psi)\| \int_0^\tau \omega(\|\psi_t - \psi\|)\mathrm{d}t \\
&\leqslant \|v(\psi)\| \, \tau \, \omega(\tau \|v(\psi)\|). \qquad (4.68)
\end{aligned}$$

To establish the first inequality, we used that ψ and ψ_t belong to the compact set K and for the second one that ω is increasing and that $t \in [0,\tau]$.

Linear convergence. We first show the existence of $\tau_1^* > 0$ such that for all $\psi \in \mathcal{S}$ and $\tau \in (0, \tau_1^*)$, one has $\psi_\tau \in \mathcal{S}$. By definition of ε, for every $i \in \{1, \dots, N\}$ one has $\nu_i \geqslant 2\epsilon$ and $G_i(\psi) \geqslant \epsilon$. Using (4.67) and (4.68), one deduces a lower bound on $G_i(\psi_\tau)$:

$$\begin{aligned}
G_i(\psi_\tau) &\geqslant (1-\tau)G_i(\psi) + \tau\nu_i + R_i(\tau) \\
&\geqslant (1+\tau)\epsilon - \|R(\tau)\| \\
&\geqslant \varepsilon + \tau(\varepsilon - M\omega(\tau M)).
\end{aligned}$$

If we choose $\tau_1^* > 0$ small enough so that $M\omega(\tau_1^* M) \leqslant \varepsilon$, this implies that $\psi_\tau \in \mathcal{S}$ for all $\psi \in \mathcal{S}$ and $\tau \in [0, \tau_1^*]$.

We now prove that there exists $\tau_2^* > 0$ such that for $\tau \in [0, \tau_2^*]$, one has $\|G(\psi_\tau) - \nu\| \leqslant (1 - \tau/2) \|G(\psi) - \nu\|$. From Equation (4.67), we have $G(\psi_\tau) - \nu = (1-\tau)(G(\psi) - \nu) + R(\tau)$, and it is therefore sufficient to prove

$$\|R(\tau)\| \leqslant \frac{\tau}{2} \|G(\psi) - \nu\|.$$

With the upper bound on $R(\tau)$ given in Equation (4.68) combined with the two bounds on $\|v(\psi)\|$ of Equation (4.66), this condition will hold provided that τ is such that $\omega(\tau M)/a \leqslant 1/2$.

These two bounds directly imply that the $\tau^{(k)}$ chosen in Algorithm 4 always satisfy $\tau^{(k)} \geqslant \tau^*$ with $\tau^* = \frac{1}{2}\min(\tau_1^*, \tau_2^*)$, so that

$$\left\| G(\psi^{(k+1)}) - \nu \right\| \leqslant \left(1 - \frac{\tau^*}{2}\right) \left\| G(\psi^{(k+1)}) - \nu \right\|.$$

This establishes the linear convergence of Algorithm 4. □

Application to optimal transport. The damped Newton algorithm allows to solve the semi-discrete optimal transport problem when applied to the function G given by $G_y(\psi) = \rho(\mathrm{Lag}_\psi(y))$. The function G satisfies the assumptions of Proposition 50: it is of class C^1 by Theorem 45, satisfies the compactness and strict monotonicity property by Theorem 47 and clearly also satisfies Assumption *(1)*. We therefore have the following theorem:

Theorem 51. *We make the following assumptions:*

- $c \in \mathcal{C}^2(\Omega_X \times \Omega_Y)$ *satisfies the twist condition (Def. 8),*
- Y *is generic with respect to* c *and* ∂X *(Def. 18),*
- $\rho(\partial X) = 0$ *and* $\rho_{|X} \in \mathcal{C}^0(X)$ *is such that the set* $\{\rho > 0\} \cap \mathrm{int}(X)$ *is connected.*

Then Algorithm 4 terminates in a finite number of steps. More precisely, the iterates $(\psi^{(k)})$ *of Algorithm 4 satisfy, for some* $\tau^* > 0$,

$$\left\|G(\psi^{k+1}) - \nu\right\| \leqslant \left(1 - \frac{\tau^\star}{2}\right) \left\|G(\psi^k) - \nu\right\|.$$

Remark 28 (Quadratic convergence). The above theorem shows that the convergence of the damped Newton algorithm is globally linear. When the cost c satisfies the *Ma-Trudinger-Wang* (MTW) condition that appears in the regularity theory of optimal transport, and when the density function ρ is Lipschitz-continuous, the convergence is even locally quadratic [65].

Remark 29 (Implementation). The most difficult part in the implementation of both Oliker–Prussner's algorithm and the damped Newton algorithm is the computation of the Laguerre tessellation. In several interesting cases, Laguerre cells can be obtained by intersecting Power diagram with surfaces, such as planes, spheres or triangulated surfaces. Recall that the Power diagram of a weighted point cloud $P = (p_1, \dots, p_N) \in (\mathbb{R}^d)^N$ with weights $(\omega_1, \dots, \omega_N) \in \mathbb{R}$, is defined by the cells

$$\mathrm{Pow}_i := \{x \in \mathbb{R}^3 \mid \|x - p_i\|^2 + \omega_i \leqslant \|x - p_j\|^2 + \omega_j \quad \forall j\}.$$

This diagram can be efficiently computed by using libraries, such as for instance CGAL or GEOGRAM.

When $c(x, y) = \|x - y\|^2$ is the quadratic cost and X is a triangulated surface in $\mathbb{R}^3$, the Laguerre cells can be obtained by intersecting power cells with the triangulated surface X [72]. This approach is also used in several inverse problems arising in nonimaging optics that correspond to optimal transport problems. For instance, when $c(x, y) = -\log(1 - \langle x|y\rangle)$ is the reflector cost on the unit sphere, the Laguerre cells are obtained by intersecting power cells with the unit sphere [74, 37].

4.4. **Semi-discrete entropic transport.** The semi-discrete entropic transport problem was introduced by Genevay, Cuturi, Peyré and Bach [52], as a regularization of high-dimensional optimal transport problems, see also [36]. Such high-dimensional problems occur for instance in image generation, we refer for instance to the work of Galerne, Leclaire and Rabin [48]. Our goal here is to investigate briefly the relation between the semi-discrete Kantorovich functional $\mathcal{K}$ and its entropically regularized variant $\mathcal{K}^\eta$.

Let $X \subseteq \Omega_X$ be compact and $Y \subseteq \Omega_Y$ be finite. We recall that the entropy of a probability measure is

$$\mathcal{H}(\rho) = \begin{cases} \int_{\Omega_X} \rho \log \rho & \text{if } \rho \in \mathcal{P}^{\mathrm{ac}}(X) \\ +\infty & \text{if not} \end{cases} \tag{4.69}$$

We also recall that if γ is a transport plan between a probability density ρ in $\mathcal{P}^{\mathrm{ac}}(X)$ and a finitely supported measure $\nu = \sum_{y \in Y} \nu_y \delta_y$, then there exists probability densities $\rho_y \in \mathcal{M}^+(X) \cap \mathrm{L}^1(X)$ such that $\gamma = \sum_y \rho_y \otimes \delta_y$, and which satisfy the two marginal conditions

$$\sum_y \rho_y = \rho \qquad \text{and} \qquad \int \rho_y = \nu_y. \tag{4.70}$$

Then, the entropy of γ, with respect to $\mathrm{vol}^d \otimes \mathrm{vol}^0$, is the sum of the entropies of the ρ_y. This leads to the following definition.

Definition 20 (Semi-discrete entropic transport)**.** The entropy-regularized semi-discrete optimal transport problem between a density $\rho \in \mathcal{P}^{\mathrm{ac}}(X)$ and a finitely supported measure $\nu = \sum_{y\in Y} \nu_y \delta_y$ is defined for any $\eta > 0$ by

$$(\mathrm{KP})^\eta = \min \left\{ \langle c|\gamma\rangle + \eta \sum_{y\in Y} \mathcal{H}(\rho_y) \mid \rho_y \in \mathrm{L}^1(X), \text{s.t. } \sum_{y\in Y} \rho_y = \rho, \int_X \rho_y = \nu_y \right\}.$$

Dual problem. The dual problem is constructed, as always, by introducing Lagrange multipliers φ, ψ for the marginal constraints (4.70). We skip the derivation of the dual problem, which is very similar to the one presented in the discrete case (Section 3.3), and we directly state it:

$$(\mathrm{DP})^\eta = \sup_{(\varphi,\psi)\in \mathrm{L}^1(X)\times\mathbb{R}^Y} \langle \varphi|\rho\rangle - \langle \psi|\nu\rangle - \eta \sum_{y\in Y} \int_X e^{-\frac{c(x,y)+\psi(y)-\varphi(x)}{\eta}} \mathrm{d}x. \tag{4.71}$$

Maximizing with respect to φ for a given $\psi \in \mathbb{R}^Y$, we obtain a second formulation as a finite-dimensional optimization problem involving a regularized Kantorovich functional, exactly as in §3.3:

$$(\mathrm{DP})^{\eta'} = \sup_{\psi\in\mathbb{R}^Y} \mathcal{K}^\eta(\psi) \tag{4.72}$$

$$\text{where } \mathcal{K}^\eta(\psi) := -\eta \int_X \log\left(\sum_{y\in Y} e^{-\frac{c(x,y)+\psi(y)}{\eta}} \right) \rho(x)\mathrm{d}x - \langle \psi|\nu\rangle + \eta\mathcal{H}(\rho)$$

In order to express the gradient and the Hessian of $\mathcal{K}^\eta$, we also need the notion of smoothed laguerre cells, introduced in the discrete case (see Equation (3.46)) and defined by

$$\mathrm{RLag}_y^\eta(\psi) = \frac{e^{-\frac{c(\cdot,y)+\psi(y)}{\eta}}}{\sum_{z\in Y} e^{-\frac{c(\cdot,z)+\psi(z)}{\eta}}}. \tag{4.73}$$

Theorem 52. *Assume that $c \in \mathcal{C}^1(X \times Y)$ is twisted. Then, $\mathcal{K}^\eta$ is a $\mathcal{C}^2$ strictly concave function over $\mathbb{R}^Y$, with first-order partial derivatives*

$$\forall y \in Y,\ \frac{\partial \mathcal{K}^\eta}{\partial \mathbf{1}_y}(\psi) = G^\eta_y(\psi) - \nu_y \text{ with } G^\eta_y(\psi) := \langle \mathrm{RLag}^\eta_y(\psi) | \rho \rangle. \tag{4.74}$$

and second-order partial derivatives

$$\begin{aligned} &\forall y \neq z \in Y,\ \frac{\partial^2 \mathcal{K}^\eta}{\partial \mathbf{1}_z \partial \mathbf{1}_y}(\psi) = G^\eta_{yz}(\psi) := \frac{1}{\eta} \langle \mathrm{RLag}^\eta_y(\psi) \mathrm{RLag}^\eta_z(\psi) | \rho \rangle \\ &\forall y \in Y,\ \frac{\partial^2 \mathcal{K}^\eta}{\partial \mathbf{1}_y^2}(\psi) = G^\eta_{yy}(\psi) := -\sum_{z \neq y} G^\eta_{yz}(\psi) \end{aligned} \tag{4.75}$$

If ψ is a maximizer in $(\mathrm{DP})^{\eta'}$, then the solution to $(\mathrm{KP})^\eta$ is given by

$$\gamma = \sum_y \rho_y \otimes \delta_y, \text{ with } \rho_y = \mathrm{RLag}^\eta_y(\psi)\rho.$$

We skip the proof of this theorem which follows closely the one of Theorem 32 in the discrete case.

Strong convergence of $\mathcal{K}^\eta$ to $\mathcal{K}$. The next proposition show that for twisted costs, G^η converges to G locally uniformly (i.e. $\mathcal{K}^\eta$ converges to $\mathcal{K}$ in $\mathcal{C}^1$). Its proof follows closely the proof of the Lipschitz estimate for G^η in Proposition 41.

Proposition 53. *Assume that $c \in \mathcal{C}^2(\Omega_X \times \Omega_Y)$ is twisted (Def 8), that $X \subseteq \Omega_X$ is compact and $Y \subseteq \Omega_Y$ is finite and that $\rho \in \mathcal{P}^{\mathrm{ac}}(X) \cap \mathrm{L}^\infty(X)$. Then:*

(i) G^η converges pointwise to G as $\eta \to 0$, i.e.

$$\forall y \in Y,\ \mathrm{RLag}^\eta_y(\psi) \xrightarrow[\mathrm{L}^1(X)]{\eta \to 0} \mathbf{1}_{\mathrm{Lag}_y(\psi)}.$$

(ii) G^η is L-Lipschitz, where L depends on c, X and N only.
(iii) G^η converges locally uniformly to G.

Remark 30. The formula (4.75) implies that G^η is $\frac{1}{\eta}$-Lipschitz continous, see [52] or Remark 5.1 in [83]. When the cost is twisted, the previous proposition shows that the family of functions $(G^\eta)_{\eta > 0}$ is in fact uniformly Lipschitz.

Proof. (i) To prove this statement, it suffices to remark that

$$\lim_{\eta \to 0} \mathrm{RLag}^\eta_y(\psi)(x) = \begin{cases} 1 \text{ if } x \in \mathrm{SLag}_y(\psi), \\ 0 \text{ if } x \in X \setminus \mathrm{Lag}_y(\psi). \end{cases}$$

Thus, $\mathrm{RLag}^\eta_y(\psi)$ converges to $\mathbf{1}_{\mathrm{Lag}_y(\psi)}$ pointwise on the complement in X of $\mathrm{Lag}_y(\psi) \setminus \mathrm{SLag}_y(\psi)$. Since the cost is twisted, $\mathrm{Lag}_y(\psi) \setminus \mathrm{SLag}_y(\psi)$ is Lebesgue-negligible, therefore proving that $\mathrm{RLag}^\eta_y(\psi)$ converges almost everywhere to $\mathbf{1}_{\mathrm{Lag}_y(\psi)}$. One concludes by applying Lebesgue's dominated convergence theorem.

(ii) To prove that G^η is Lipschitz, we compute an upper bound on $\mathrm{D}G^\eta = \mathrm{D}^2\mathcal{K}^\eta$, recalling that for $z \neq y$,

$$\frac{\partial^2 \mathcal{K}^\eta}{\partial \mathbf{1}_z \partial \mathbf{1}_y}(\psi) = \frac{1}{\eta} \langle \mathrm{RLag}^\eta_y(\psi) \mathrm{RLag}^\eta_z(\psi) | \rho \rangle.$$

For getting such an upper bound, as in Proposition 41, we will apply the co-area formula using the function

$$f(x) = c(x,z) + \psi(z) - (c(x,y) + \psi(y))$$

We note that

$$\mathrm{RLag}_y^\eta(\psi)(x)\mathrm{RLag}_z^\eta(\psi)(x) = \frac{e^{-\frac{c(x,y)+\psi(y)+c(x,z)+\psi(z)}{\eta}}}{\left(\sum_{w\in Y} e^{-\frac{c(x,w)+\psi(w)}{\eta}}\right)^2}.$$

When $f(x) \geqslant 0$, we use the equality $c(x,z) + \psi(z) = c(x,y) + \psi(y) + f(x)$ to obtain the upper bound

$$\frac{e^{-\frac{c(x,y)+\psi(y)+c(x,z)+\psi(z)}{\eta}}}{\left(\sum_{w\in Y} e^{-\frac{c(x,w)+\psi(w)}{\eta}}\right)^2} \leqslant \frac{\left(e^{-\frac{c(x,y)+\psi(y)}{\eta}}\right)^2 e^{-\frac{f(x)}{\eta}}}{\left(\sum_{w\in Y} e^{-\frac{c(x,w)+\psi(w)}{\eta}}\right)^2} \leqslant e^{-\frac{f(x)}{\eta}}.$$

Reasoning similarly when $f(x) \leqslant 0$, we obtain

$$\mathrm{RLag}_y^\eta(\psi)(x)\mathrm{RLag}_z^\eta(\psi)(x) \leqslant e^{-\frac{|f(x)|}{\eta}}.$$

Using the co-area formula and the previous upper bound, we get

$$\begin{aligned}
&\langle \mathrm{RLag}_y^\eta(\psi)\mathrm{RLag}_z^\eta(\psi)|\rho\rangle \\
&= \int_X \mathrm{RLag}_y^\eta(\psi)(x)\mathrm{RLag}_z^\eta(\psi)(x)\rho(x)\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\int_{f^{-1}(t)} \mathrm{RLag}_y^\eta(\psi)(x)\mathrm{RLag}_z^\eta(\psi)(x)\frac{\rho(x)}{\|\nabla f(x)\|}\mathrm{dvol}^{d-1}(x)\mathrm{d}t \\
&\leqslant \int_{-\infty}^{\infty}\int_{f^{-1}(t)\cap X} e^{-\frac{|t|}{\eta}}\frac{\rho(x)}{\|\nabla f(x)\|}\mathrm{dvol}^{d-1}(x)\mathrm{d}t
\end{aligned}$$

We now apply Lemma 43, which gives an upper bound on $\mathrm{vol}^{d-1}(f^{-1}(t)\cap X)$ in terms of the constants $\kappa_{yz} = \min_X \|\nabla f\|$ and $M_{yz} = \max_X \left\|\mathrm{D}^2 f\right\|$:

$$\begin{aligned}
&\langle \mathrm{RLag}_y^\eta(\psi)\mathrm{RLag}_z^\eta(\psi)|\rho\rangle \\
&\leqslant c(d)\frac{\|\rho\|_\infty}{\kappa_{yz}}\left(1 + \frac{M_{yz}}{\kappa_{yz}}\mathrm{diam}(X)\right)\mathrm{diam}(X)^{d-1}\int_{-\infty}^{\infty} e^{-\frac{|t|}{\eta}}\mathrm{d}t \\
&\leqslant C\eta,
\end{aligned}$$

where the constant C depends on the domain, ρ and the cost only. In other words, for $z \neq y$,

$$\left|\frac{\partial^2\mathcal{K}^\eta}{\partial\mathbf{1}_z\partial\mathbf{1}_y}(\psi)\right| = \frac{1}{\eta}\langle \mathrm{RLag}_y^\eta(\psi)\mathrm{RLag}_z^\eta(\psi)|\rho\rangle \leqslant C.$$

A similar upper bound holds Since the diagonal elements, thus ensuring that G^η is L-Lipschitz with L independent on η. (iii) follows at once from pointwise convergence and the uniform Lipschitz estimate. □

To finish this section, we show that under the genericity assumption introduced in Section 4.3, the Hessian of Kantorovich's regularized functional $\mathrm{D}^2\mathcal{K}^\eta$ converges pointwise to $\mathrm{D}^2\mathcal{K}$ as η converges to 0.

Theorem 54. *Assume that $c \in \mathcal{C}^1(X \times Y)$ is twisted (Def 8), that Y is generic with respect to c and ∂X (Def 18), that $\rho_{|X} \in \mathcal{C}^0(X)$. Then,*

$$\forall \psi \in \mathbb{R}^Y, \ \lim_{\eta \to 0} \mathrm{D}G^\eta(\psi) = \mathrm{D}G(\psi).$$

Proof. We let $f(x) = c(x,z) + \psi(z) - (c(x,y) + \psi(y))$ and $\epsilon > 0$. From the proof of Proposition 53, one has for every $x \in X \setminus f^{-1}([-\varepsilon, \varepsilon])$ that

$$\mathrm{RLag}_y^\eta(\psi)(x)\mathrm{RLag}_z^\eta(\psi)(x) \leqslant e^{-\frac{\varepsilon}{\eta}}.$$

This implies that

$$\lim_{\eta \to 0} \int_{X \setminus f^{-1}([-\varepsilon,\varepsilon])} \mathrm{RLag}_y^\eta(\psi)(x)\mathrm{RLag}_z^\eta(\psi)(x)\rho(x)\mathrm{d}x = 0.$$

Since the smoothed Laguerre are non-negative on X, we get

$$G_{yz}^\eta(\psi) = \frac{1}{\eta}\int_X \mathrm{RLag}_y^\eta(\psi)(x)\mathrm{RLag}_z^\eta(\psi)(x)\rho(x)\mathrm{d}x \underset{\eta \to 0}{\sim} G_{yz}^{\eta,\varepsilon}(\psi)$$

where

$$G_{yz}^{\eta,\varepsilon}(\psi) := \frac{1}{\eta}\int_{X \cap f^{-1}([-\varepsilon,\varepsilon])} \mathrm{RLag}_y^\eta(\psi)(x)\mathrm{RLag}_z^\eta(\psi)(x)\rho(x)\mathrm{d}x.$$

As in the proof of Lemma 46, we construct $\Phi : [-\varepsilon, \varepsilon] \times M \to \Omega_X$ (for some positive ε), with $M = f^{-1}(0)$ by solving the Cauchy problem

$$\begin{cases} \Phi(0,x) = x \\ \frac{d}{dt}\Phi(t,x) = \frac{\nabla f(\Phi(t,x))}{\|\nabla f(\Phi(t,x))\|^2}, \end{cases}$$

so that $f(\Phi(t, f^{-1}(0))) = t$. Then one has $\Phi([-\epsilon,\epsilon] \times M) = X \cap f^{-1}([-\epsilon,\epsilon])$, and by a change of variable formula, using the definition of the smoothed indicator function of Laguerre cells and the definition of f, we get

$$\begin{aligned} &G_{yz}^{\eta,\varepsilon}(\psi) \\ =&\frac{1}{\eta}\int_M \int_{-\varepsilon}^{\varepsilon} \mathrm{RLag}_y^\eta(\psi)(\Phi(t,x))\mathrm{RLag}_z^\eta(\psi)(\Phi(t,x))\rho(\Phi(t,x))J_\Phi(t,x)\mathrm{d}t\mathrm{dvol}^{d-1}(x). \end{aligned}$$

Remark that

$$\begin{aligned} \mathrm{RLag}_y^\eta(\psi)(\Phi(t,x))\mathrm{RLag}_z^\eta(\psi)(\Phi(t,x)) &= \frac{e^{-\frac{c(\Phi(t,x),y)+\psi(y)+c(\Phi(t,x),z)+\psi(z)}{\eta}}}{\left(\sum_{z\in Y} e^{-\frac{c(\Phi(t,x),z)+\psi(z)}{\eta}}\right)^2} \\ &= \chi_\eta(t,x)e^{-\frac{|f(\Phi(t,x))|}{\eta}} \\ &= \chi_\eta(t,x)e^{-\frac{|t|}{\eta}}, \end{aligned}$$

where we put

$$\chi_\eta(t,x) := \frac{e^{-2\min\left(\frac{c(\Phi(t,x),y)+\psi(y)}{\eta}, \frac{c(\Phi(t,x),z)+\psi(z)}{\eta}\right)}}{\left(\sum_{z\in Y} e^{-\frac{c(\Phi(t,x),z)+\psi(z)}{\eta}}\right)^2}.$$

We deduce that one gets

$$G_{yz}^{\eta,\varepsilon}(\psi) = \int_M g_\eta(x) \mathrm{dvol}^{d-1}(x)$$

with

$$g_\eta(x) = \int_{-\varepsilon}^{\varepsilon} \chi_\eta(t,x) \frac{e^{-\frac{|t|}{\eta}}}{\eta} \rho(\Phi(t,x)) J_\Phi(t,x) \mathrm{d}t$$

We first note that $|\chi_\eta(t,x)| \leqslant 1$, so that $(g_\eta)_\eta$ is bounded in $\mathrm{L}^1(M)$. We now prove that $g_\eta(x)$ converges to $g(x) = \rho(x) J_\Phi(0,x) \mathbf{1}_{\mathrm{Lag}_{yz}(\psi)}(x)$ vol^{d-1}-almost everywhere. More precisely, we show convergence for any x belonging to the following set E, which has full vol^{d-1} measure in $\mathrm{Lag}_{yz}(\psi)$ by the genericity assumption (Def. 18):

$$E = \mathrm{Lag}_{yz}(\psi) \setminus \left(\partial X \cup \bigcup_{w \notin \{y,z\}} \mathrm{Lag}_w(\psi) \right).$$

We split the integral defining g_η by distinguishing the case $t \leqslant 0$ and $t \geqslant 0$. For $t \geqslant 0$,

$$t = f(\Phi(t,x)) = c(\Phi(x,t), z) + \psi(z) - (c(\Phi(x,t), y) + \psi(y)) \geqslant 0$$

giving

$$\begin{aligned}
\chi_\eta(t,x) &= \frac{\left(e^{-\frac{c(\Phi(t,x),y)+\psi(y)}{\eta}} \right)^2}{\left(\sum_{w \in Y} e^{-\frac{c(\Phi(t,x),w)+\psi(w)}{\eta}} \right)^2} \\
&= \left(\sum_{w \in Y} e^{-\frac{c(\Phi(t,x),w)+\psi(w)-(c(\Phi(t,x),y)+\psi(y))}{\eta}} \right)^{-2} \\
&= \left(1 + e^{-\frac{t}{\eta}} + r_\eta(t,x) \right)^{-2}
\end{aligned}$$

with

$$r_\eta(t,x) = \sum_{w \in Y \setminus \{y,z\}} e^{-\frac{1}{\eta}(c(\Phi(t,x),w)+\psi(w)-(c(\Phi(t,x),y)+\psi(y))}.$$

Now, by assumption on the point x, for any $w \notin \{y,z\}$, one has $c(x,y) + \psi(y) < c(x,w) + \psi(w)$ so that $r_\eta(t,x)$ is negligible A similar computation can be done for $t \leqslant 0$, giving us the estimation

$$g_\eta(x) \underset{\eta \to 0}{\sim} \int_{-\varepsilon}^{\varepsilon} \frac{e^{-\frac{|t|}{\eta}}}{\eta(1 + e^{-\frac{|t|}{\eta}})^2} \rho(\Phi(t,x)) J_\Phi(t,x) \mathrm{d}t \xrightarrow{\eta \to 0} \rho(x) J_\Phi(0,x).$$

On the other hand, one can show that for almost every x in M but not in $\mathrm{Lag}_{yz}(\psi)$, $|\chi_\eta(t,x)|$ tends to zero when η goes to zero, thus implying that the sequence $(g_\eta(x))$ also converges to 0. In other words,

$$g_\eta(x) \xrightarrow[a.e.]{\eta \to 0} \rho(x) J_\Phi(0,x) \mathbf{1}_{\mathrm{Lag}_{yz}(\psi)}(x) = \frac{\rho(x)}{\|\nabla f(x)\|} \mathbf{1}_{\mathrm{Lag}_{yz}(\psi)}(x).$$

By Lebesgue's dominated convergence theorem, we get

$$\lim_{\eta\to 0} G^{\eta}_{yz}(\psi) = \lim_{\eta\to 0} \int_M g_\eta(x) \mathrm{dvol}^{d-1}(x) = \int_{\mathrm{Lag}_{yz}(\psi)} \frac{\rho(x)}{\|\nabla f(x)\|} \mathrm{dvol}^{d-1}(x).$$

From the relation $\|\nabla f(x)\| = \|\nabla_x c(x,y) - \nabla_x c(x,z)\|$ we get as desired,

$$\lim_{\eta\to 0} G^{\eta}_{yz}(\psi) = G_{yz}(\psi) \qquad \square$$

5. Appendix

5.1. **Convex analysis.** We recall a few relevant definitions and facts from convex analysis (adapted to concave functions).

Definition 21. The superdifferential of function $F : \mathbb{R}^N \to \mathbb{R} \cup \{-\infty\}$ at $x \in \mathbb{R}^N$ is the set of vectors $v \in \mathbb{R}^N$ such that

$$\forall y \in Y, \quad F(y) \leqslant F(x) + \langle v | y - x \rangle.$$

This set is denoted $\partial^+ F(x)$.

Proposition 55. *The following hold:*

- *A function $F : \mathbb{R}^N \to \mathbb{R}$ is concave if and only if*

$$\forall x \in \mathbb{R}^N, \partial^+ F(x) \neq \emptyset.$$

- *The superdifferential can be characterized by (*[86, Theorem 25.6]*):*

$$\partial^+ F(x) = \mathrm{conv}\left\{ \lim_{n\to\infty} \nabla F(x_n) \mid (x_n)_{n\in\mathbb{N}} \in S \right\}, \tag{5.76}$$

where $\mathrm{conv}(Z)$ *denotes the convex envelope of the set Z and*

$$S = \{(x_n)_{n\in\mathbb{N}} \mid \forall n \geqslant 1, \nabla F(x_n) \text{ exists and } \nabla F(x_n) \text{ exists}\}.$$

5.2. **Coarea formula.** We consider two Riemannian sub-manifolds M and N, respectively of dimensions m and n, of two Euclidean spaces and assume that $n \leqslant m$. Let $\Phi : M \to N$ be a function of class C^1 between the two manifolds. The *Jacobian determinant* of $\Phi : E \subseteq M \to N$ at x is defined by

$$J_\Phi(x) = \sqrt{\det(\mathrm{D}\Phi(x)\mathrm{D}\Phi(x)^T)} \tag{5.77}$$

Note that if $M = \mathbb{R}^m$, $N = \mathbb{R}^n$ and $\Phi = (\Phi_1, \dots, \Phi_n)$, one has

$$\mathrm{D}\Phi(x)\mathrm{D}\Phi(x)^T = (\langle \nabla\Phi_i(x) | \nabla\Phi_j(x) \rangle)_{1\leqslant i,j\leqslant n}.$$

In particular, for $n = 1$, one has $J\Phi(x) = \|\nabla\Phi_1(x)\|$, and for $n = 2$ one gets

$$J\Phi(x)^2 = \|\nabla\Phi_1(x)\|^2 \|\nabla\Phi_2(x)\|^2 - \langle \nabla\Phi_1(x) | \nabla\Phi_2(x) \rangle^2$$

which by Cauchy-Schwarz's inequality is always non-negative and vanishes iff $\nabla\Phi_1(x)$ and $\nabla\Phi_2(x)$ are collinear.

Theorem 56 (Coarea formula). *Let $\Phi : M \to N$ be a function of class C^1. For every vol^m-measurable function $u : M \to \mathbb{R}$, one has*

$$\int_M u(x) J_\Phi(x) \mathrm{dvol}^m(x) = \int_N \int_{\Phi^{-1}(y)} u(x) \mathrm{dvol}^{m-n}(x) \mathrm{dvol}^n(y).$$

If $J_\Phi(x)$ does not vanish on a measurable subset $E \subset M$, then

$$\int_M u(x) \mathrm{dvol}^m(x) = \int_N \int_{\Phi^{-1}(y)} \frac{u(x)}{J_\Phi(x)} \mathrm{dvol}^{m-n}(x) \mathrm{dvol}^n(y). \tag{5.78}$$

In particular, if $m = n$ and $\Phi : M \to N$ is an homeomorphism of class C^1, letting $v = u \circ \Phi^{-1} : N \to \mathbb{R}$, one recovers the change of variable formula

$$\int_M v(\Phi(x)) J_\Phi(x) \mathrm{dvol}^m(x) = \int_N v(y) \mathrm{dvol}^n(y). \tag{5.79}$$

References

1. Pankaj K Agarwal and R Sharathkumar, *Approximation algorithms for bipartite matching with metric and geometric costs*, Proceedings of the forty-sixth annual ACM symposium on Theory of computing, ACM, 2014, pp. 555–564.
2. Martial Agueh and Guillaume Carlier, *Barycenters in the wasserstein space*, SIAM Journal on Mathematical Analysis **43** (2011), no. 2, 904–924.
3. Jason Altschuler, Jonathan Weed, and Philippe Rigollet, *Near-linear time approximation algorithms for optimal transport via sinkhorn iteration*, Advances in Neural Information Processing Systems, 2017, pp. 1964–1974.
4. Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.
5. Luigi Ambrosio, Federico Glaudo, and Dario Trevisan, *On the optimal map in the 2-dimensional random matching problem*, arXiv preprint arXiv:1903.12153, 2019.
6. Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov, *Minkowski-type theorems and least-squares clustering*, Algorithmica **20** (1998), no. 1, 61–76.
7. J-D Benamou, Yann Brenier, and Kevin Guittet, *The monge–kantorovitch mass transfer and its computational fluid mechanics formulation*, International Journal for Numerical methods in fluids **40** (2002), no. 1-2, 21–30.
8. Jean-David Benamou and Yann Brenier, *Mixed* L^2*-Wasserstein optimal mapping between prescribed density functions*, Journal of Optimization Theory and Applications **111** (2001), no. 2, 255–271.
9. Jean-David Benamou and Guillaume Carlier, *Augmented lagrangian methods for transport optimization, mean field games and degenerate elliptic equations*, Journal of Optimization Theory and Applications **167** (2015), no. 1, 1–26.
10. Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré, *Iterative bregman projections for regularized transportation problems*, SIAM Journal on Scientific Computing **37** (2015), no. 2, A1111–A1138.
11. Jean-David Benamou, Guillaume Carlier, and Luca Nenna, *A numerical method to solve multi-marginal optimal transport problems with coulomb cost*, Splitting Methods in Communication, Imaging, Science, and Engineering, Springer, 2016, pp. 577–601.
12. Jean-David Benamou and Vincent Duval, *Minimal convex extensions and finite difference discretisation of the quadratic monge–kantorovich problem*, European Journal of Applied Mathematics **30** (2019), no. 6, 1041–1078.
13. Jean-David Benamou, Brittany D Froese, and Adam M Oberman, *Numerical solution of the optimal transportation problem using the monge–ampère equation*, Journal of Computational Physics **260** (2014), 107–126.
14. Robert J. Berman, *The Sinkhorn algorithm, parabolic optimal transport and geometric Monge-Ampère equations*, arXiv preprint arXiv:1712.03082, 2017.
15. Robert J Berman, *Convergence rates for discretized monge-ampére equations and quantitative stability of optimal transport*, arXiv preprint arXiv:1803.00785, 2018.

16. D.P. Bertsekas, *A new algorithm for the assignment problem*, Mathematical Programming **21** (1981), no. 1, 152–171.
17. D.P. Bertsekas and J. Eckstein, *Dual coordinate step methods for linear network flow problems*, Mathematical Programming **42** (1988), no. 1, 203–243.
18. Garrett Birkhoff, *Tres observaciones sobre el algebra lineal*, Univ. Nac. Tucuman, Ser. A **5** (1946), 147–154.
19. Yann Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Communications on pure and applied mathematics **44** (1991), no. 4, 375–417.
20. ______, *Minimal geodesics on groups of volume-preserving maps and generalized solutions of the euler equations*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences **52** (1999), no. 4, 411–452.
21. R.E. Burkard, M. Dell'Amico, and S. Martello, *Assignment problems*, Society for Industrial Mathematics, 2009.
22. Giuseppe Buttazzo, Luigi De Pascale, and Paola Gori-Giorgi, *Optimal-transport formulation of electronic density-functional theory*, Physical Review A **85** (2012), no. 6, 062502.
23. Giuseppe Buttazzo, Chloé Jimenez, and Edouard Oudet, *An optimization problem for mass transportation with congested dynamics*, SIAM Journal on Control and Optimization **48** (2009), no. 3, 1961–1976.
24. LA Caffarelli and VI Oliker, *Weak solutions of one inverse problem in geometric optics*, Journal of Mathematical Sciences **154** (2008), no. 1, 39–49.
25. Luis Caffarelli and Robert J McCann, *Free boundaries in optimal transport and monge-ampere obstacle problems*, Annals of mathematics **171** (2010), no. 2, 673–730.
26. Luis A Caffarelli, Sergey A Kochengin, and Vladimir I Oliker, *Problem of reflector design with given far-field scattering data*, Monge Ampère Equation: Applications to Geometry and Optimization: NSF-CBMS Conference on the Monge Ampère Equation, Applications to Geometry and Optimization, July 9-13, 1997, Florida Atlantic University, vol. 226, American Mathematical Soc., 1999, p. 13.
27. Guillaume Carlier, Victor Chernozhukov, Alfred Galichon, et al., *Vector quantile regression: an optimal transport approach*, The Annals of Statistics **44** (2016), no. 3, 1165–1192.
28. Jose A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei, *Primal dual methods for wasserstein gradient flows*, arXiv preprint arXiv:1901.08081, 2019.
29. Benjamin Charlier, Jean Feydy, Joan Alexis Glaunes, and Alain Trouvé, *An efficient kernel product for automatic differentiation libraries, with applications to measure transport*, Working version, 2017.
30. Victor Chernozhukov, Alfred Galichon, Marc Hallin, Marc Henry, et al., *Monge–kantorovich depth, quantiles, ranks and signs*, The Annals of Statistics **45** (2017), no. 1, 223–256.
31. Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, *An interpolating distance between optimal transport and fisher–rao metrics*, Foundations of Computational Mathematics **18** (2018), no. 1, 1–44.
32. Codina Cotar, Gero Friesecke, and Claudia Klüppelberg, *Density functional theory and optimal transportation with coulomb cost*, Communications on Pure and Applied Mathematics **66** (2013), no. 4, 548–599.
33. Keenan Crane, Clarisse Weischedel, and Max Wardetzky, *Geodesics in heat: A new approach to computing distance based on heat flow*, ACM Transactions on Graphics (TOG) **32** (2013), no. 5, 152.
34. Michael JP Cullen and R James Purser, *An extended lagrangian theory of semi-geostrophic frontogenesis*, Journal of the atmospheric sciences **41** (1984), no. 9, 1477–1497.
35. Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems, 2013, pp. 2292–2300.
36. Marco Cuturi and Gabriel Peyré, *Semidual regularized optimal transport*, SIAM Review **60** (2018), no. 4, 941–965.

37. Pedro Machado Manhães de Castro, Quentin Mérigot, and Boris Thibert, *Far-field reflector problem and intersection of paraboloids*, Numerische Mathematik **134** (2016), no. 2, 389–411.
38. Fernando De Goes, Katherine Breeden, Victor Ostromoukhov, and Mathieu Desbrun, *Blue noise through optimal transport*, ACM Transactions on Graphics (TOG) **31** (2012), no. 6, 171.
39. Fernando de Goes, Corentin Wallez, Jin Huang, Dmitry Pavlov, and Mathieu Desbrun, *Power particles: an incompressible fluid solver based on power diagrams.*, ACM Trans. Graph. **34** (2015), no. 4, 50–1.
40. Frédéric De Gournay, Jonas Kahn, and Léo Lebrat, *3/4-discrete optimal transport*, arXiv preprint arXiv:1806.09537, 2018.
41. ———, *Differentiation and regularity of semi-discrete optimal transport with respect to the parameters of the discrete measure*, Numerische Mathematik **141** (2019), no. 2, 429–453.
42. Roberto De Leo, Cristian E Gutiérrez, and Henok Mawi, *On the numerical solution of the far field refractor problem*, Nonlinear Analysis **157** (2017), 123–145.
43. J. Edmonds and R.M. Karp, *Theoretical improvements in algorithmic efficiency for network flow problems*, Journal of the ACM (JACM) **19** (1972), no. 2, 248–264.
44. Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon, *Computation of optimal transport on discrete metric measure spaces*, arXiv preprint arXiv:1707.06859, 2017.
45. Jean Feydy, Pierre Roussillon, Alain Trouvé, and Pietro Gori, *Fast and scalable optimal transport for brain tractograms*, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 636–644.
46. Brittany D Froese, *A numerical method for the elliptic Monge–Ampère equation with transport boundary conditions*, SIAM Journal on Scientific Computing **34** (2012), no. 3, A1432–A1459.
47. H.N. Gabow and R.E. Tarjan, *Faster scaling algorithms for network problems*, SIAM Journal on Computing **18** (1989), 1013.
48. Bruno Galerne, Arthur Leclaire, and Julien Rabin, *A texture synthesis model based on semi-discrete optimal transport in patch space*, SIAM Journal on Imaging Sciences **11** (2018), no. 4, 2456–2493.
49. Alfred Galichon, *Optimal transport methods in economics*, Princeton University Press, 2018.
50. Alfred Galichon and Bernard Salanié, *Matching with trade-offs: Revealed preferences over competing characteristics*, Tech. report, CEPR Discussion Papers, 2010.
51. Wilfrid Gangbo and Robert J McCann, *The geometry of optimal transportation*, Acta Mathematica **177** (1996), no. 2, 113–161.
52. Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach, *Stochastic optimization for large-scale optimal transport*, Advances in neural information processing systems, 2016, pp. 3440–3448.
53. Nicola Gigli, *On hölder continuity-in-time of the optimal transport map towards measures along a curve*, Proceedings of the Edinburgh Mathematical Society **54** (2011), no. 2, 401–409.
54. A.V. Goldberg, *Efficient graph algorithms for sequential and parallel computers*, Ph.D. thesis, Massachussetts Institute of Technology, 1987.
55. Xianfeng Gu, Feng Luo, Jian Sun, and Shing-Tung Yau, *Variational principles for minkowski type problems, discrete optimal transport, and discrete monge–ampère equations*, Asian Journal of Mathematics **20** (2016), no. 2, 383–398.
56. Nestor Guillen, *A primer on generated jacobian equations: Geometry, optics, economics*, Notices of the American Mathematical Society **66** (2019), no. 9.
57. Cristian E Gutiérrez and Haim Brezis, *The monge-ampere equation*, vol. 44, Springer, 2001.
58. Valentin Hartmann and Dominic Schuhmacher, *Semi-discrete optimal transport – the case* $p = 1$, arXiv preprint arXiv:1706.07650, 2017.

59. Morgane Henry, Emmanuel Maitre, and Valérie Perrier, *Optimal transport using helmholtz-hodge decomposition and first-order primal-dual algorithms*, 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 4748–4752.
60. Romain Hug, *Analyse mathématique et convergence d'un algorithme pour le transport optimal dynamique: cas des plans de transports non réguliers, ou soumis à des contraintes*, Thèse de doctorat de l'Université Grenoble-Alpes, 2016.
61. Jan-Christian Hütter and Philippe Rigollet, *Minimax rates of estimation for smooth optimal transport maps*, arXiv preprint arXiv:1905.05828, 2019.
62. Richard Jordan, David Kinderlehrer, and Felix Otto, *The variational formulation of the fokker–planck equation*, SIAM journal on mathematical analysis **29** (1998), no. 1, 1–17.
63. Michael Kerber, Dmitriy Morozov, and Arnur Nigmetov, *Geometry helps to compare persistence diagrams*, Journal of Experimental Algorithmics (JEA) **22** (2017), 1–4.
64. Jun Kitagawa, *An iterative scheme for solving the optimal transportation problem*, Calculus of Variations and Partial Differential Equations **51** (2014), no. 1-2, 243–263.
65. Jun Kitagawa, Quentin Mérigot, and Boris Thibert, *Convergence of a newton algorithm for semi-discrete optimal transport*, Journal of the European Mathematical Society (2019), OnlineFirst.
66. Stanislav Kondratyev, Léonard Monsaingeon, Dmitry Vorotnikov, et al., *A new optimal transport distance on the space of finite radon measures*, Advances in Differential Equations **21** (2016), no. 11/12, 1117–1164.
67. Hugo Lavenant, *Unconditional convergence for discretizations of dynamical optimal transport*, arXiv preprint arXiv:1909.08790, 2019.
68. Bruno Lévy, *A numerical algorithm for l2 semi-discrete optimal transport in 3d*, ESAIM: Mathematical Modelling and Numerical Analysis **49** (2015), no. 6, 1693–1715.
69. Damiano Lombardi and Emmanuel Maitre, *Eulerian models and algorithms for unbalanced optimal transport*, ESAIM: Mathematical Modelling and Numerical Analysis **49** (2015), no. 6, 1717–1744.
70. Quentin Mérigot, *A multiscale approach to optimal transport*, Computer Graphics Forum **30** (2011), no. 5, 1583–1592.
71. Quentin Mérigot, Alex Delalande, and Frédéric Chazal, *Quantitative stability of optimal transport maps and linearization of the 2-wasserstein space*, arXiv preprint arXiv:1910.05954 (2019).
72. Quentin Mérigot, Jocelyn Meyron, and Boris Thibert, *An algorithm for optimal transport between a simplex soup and a point cloud*, SIAM Journal on Imaging Sciences **11** (2018), no. 2, 1363–1389.
73. Quentin Mérigot and Jean-Marie Mirebeau, *Minimal geodesics along volume-preserving maps, through semidiscrete optimal transport*, SIAM Journal on Numerical Analysis **54** (2016), no. 6, 3465–3492.
74. Jocelyn Meyron, Quentin Mérigot, and Boris Thibert, *Light in power: a general and parameter-free algorithm for caustic design*, ACM Transactions on Graphics (TOG) **37** (2019), no. 6, 224.
75. Jean-Marie Mirebeau, *Discretization of the 3d Monge-Ampère operator, between wide stencils and power diagrams*, ESAIM: Mathematical Modelling and Numerical Analysis **49** (2015), no. 5, 1511–1523.
76. Gaspard Monge, *Mémoire sur la théorie des déblais et des remblais*, 1781.
77. Michael Neilan, Abner J Salgado, and Wujun Zhang, *The monge-amp\{e} re equation*, arXiv preprint arXiv:1901.05108, 2019.
78. Adam M Oberman and Yuanlong Ruan, *An efficient linear programming method for optimal transportation*, arXiv preprint arXiv:1509.03668, 2015.
79. VI Oliker and LD Prussner, *On the numerical solution of the equation and its discretizations, i*, Numerische Mathematik **54** (1989), no. 3, 271–293.
80. Vladimir Oliker, *Mathematical aspects of design of beam shaping surfaces in geometrical optics*, Trends in Nonlinear Analysis, Springer, 2003, pp. 193–224.
81. Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet, *Optimal transport with proximal splitting*, SIAM Journal on Imaging Sciences **7** (2014), no. 1, 212–238.

82. Brendan Pass, *Multi-marginal optimal transport: theory and applications*, ESAIM: Mathematical Modelling and Numerical Analysis **49** (2015), no. 6, 1771–1790.
83. Gabriel Peyré and Marco Cuturi, *Computational optimal transport*, Foundations and Trends® in Machine Learning **11** (2019), no. 5-6, 355–607.
84. Svetlozar T Rachev and Ludger Rüschendorf, *Mass transportation problems: Volume i: Theory*, vol. 1, Springer Science & Business Media, 1998.
85. ———, *Mass transportation problems: Applications*, Springer Science & Business Media, 2006.
86. R Tyrrell Rockafellar, *Convex analysis*, vol. 28, Princeton university press, 1970.
87. Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas, *The earth mover's distance as a metric for image retrieval*, International journal of computer vision **40** (2000), no. 2, 99–121.
88. Filippo Santambrogio, *Optimal transport for applied mathematicians*, Springer, 2015.
89. Bernhard Schmitzer, *A sparse multiscale algorithm for dense optimal transport*, Journal of Mathematical Imaging and Vision **56** (2016), no. 2, 238–259.
90. ———, *Stabilized sparse scaling algorithms for entropy regularized transport problems*, SIAM Journal on Scientific Computing **41** (2019), no. 3, A1443–A1481.
91. Richard Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, The annals of mathematical statistics **35** (1964), no. 2, 876–879.
92. Richard Sinkhorn and Paul Knopp, *Concerning nonnegative matrices and doubly stochastic matrices*, Pacific Journal of Mathematics **21** (1967), no. 2, 343–348.
93. Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas, *Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains*, ACM Transactions on Graphics (TOG) **34** (2015), no. 4, 66.
94. Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher, *Earth mover's distances on discrete surfaces*, ACM Transactions on Graphics (TOG) **33** (2014), no. 4, 67.
95. Neil S Trudinger, *On the local theory of prescribed jacobian equations*, Discrete & Continuous Dynamical Systems-A **34** (2014), no. 4, 1663–1681.
96. François-Xavier Vialard, *An elementary introduction to entropic regularization and proximal methods for numerical optimal transport*, Lecture, May 2019.
97. Cédric Villani, *Topics in optimal transportation*, no. 58, American Mathematical Soc., 2003.
98. ———, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.
99. Xu-Jia Wang, *On the design of a reflector antenna ii*, Calculus of Variations and Partial Differential Equations **20** (2004), no. 3, 329–341.